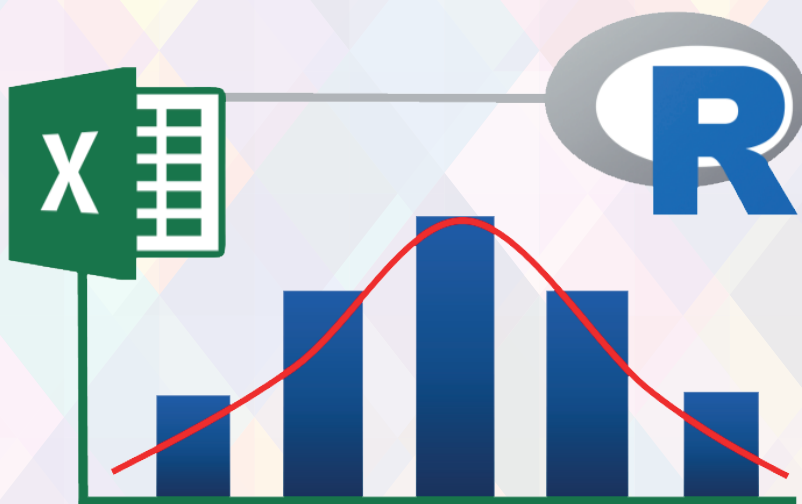


APLICACIONES DE ESTADÍSTICA BÁSICA

En MICROSOFT® EXCEL y R



*“Por un Desarrollo Agrario
Integral y Sostenible”*

*Elaborado por:
Miguel Garmendia Zapata*

En Microsoft® Excel y R

APLICACIONES DE ESTADÍSTICA BÁSICA

En MICROSOFT® EXCEL y R

Miguel Garmendia Zapata

N

005.369

G233 Garmendia Zapata, Miguel

Aplicaciones de estadística básica en
Microsoft® Excel y R / Miguel Garmendia
Zapata. -- 1a ed. -- Managua : UNA, 2020.
287 p.

ISBN 978-99924-1-044-8

1. MICROSOFT EXCEL Y R (PROGRAMA
PARA COMPUTADOR) 2. PROCESAMIENTO
ELECTRONICO DE DATOS EN ESTADISTICA

®Todos los derechos reservados 2020

©Universidad Nacional Agraria

Centro Nacional de Información y Documentación Agropecuaria

Km. 12½ Carretera Norte, Managua, Nicaragua

Teléfonos: 22331871

MSc. Miguel Garmendia Zapata

Profesor Titular, FARENA-UNA

La UNA propicia la amplia disseminación de sus publicaciones impresas y electrónicas para que el público y la sociedad en general obtenga el máximo beneficio. Por tanto en la mayoría de los casos, los colegas que trabajan en docencia, investigación y desarrollo no deben sentirse limitados en el uso de los materiales de la UNA para fines académicos y no comerciales. Sin embargo, la UNA prohíbe la modificación parcial o total de este material y espera recibir los créditos merecidos por ellos

***Esta obra está dedicada a la memoria de mis padres
Manuel Ismael Garmendia y Mariana Eugenia
Zapata, a quienes debo todo lo que actualmente soy.***

TABLA DE CONTENIDO

Prólogo.....	1
Introducción.....	2
Selección de la prueba estadística a usar	3
Clave para la selección de análisis o pruebas estadísticas.....	4
Glosario de términos relacionados con la clave.....	7
Estadísticas básicas en Microsoft Excel.....	12
Nociones sobre el ambiente de Microsoft Excel.....	12
Generalidades sobre una hoja de cálculo.....	12
Operaciones y funciones.....	14
Fórmulas en MS Excel.....	16
La herramienta de análisis de datos.....	19
Complementos para MS Excel.....	20
Estadística descriptiva en Microsoft Excel.....	20
Aplicando estadística descriptiva.....	20
La distribución normal.....	26
Evaluación de la distribución normal.....	26
Histograma.....	27
Gráfico Q.....	29
Métodos numéricos (curtosis y coeficiente de asimetría).....	36
Transformaciones.....	37
Estadística inferencial.....	38
¿Qué es “p”? ¿Probabilidad de qué? ¿Alfa? ¿Hipótesis?.....	38
Comparación de proporciones y frecuencias.....	44
Prueba de una proporción.....	45
Pruebas de dos proporciones.....	50
Prueba de bondad de ajuste.....	54
Pruebas de independencia (tablas de contingencia).....	56
Tablas de contingencia 2 x 2.....	56

Aplicaciones de Estadística Básica

Tablas de contingencia R x C.....	59
Comparación de medias.....	61
Prueba T para una muestra.....	63
Prueba T para dos muestras independientes.....	66
Prueba T para dos muestras pareadas.....	68
Análisis de varianza para un factor.....	71
Análisis de varianza para dos factores.....	73
ANDEVA de dos factores con replicación.....	74
ANDEVA de dos factores sin replicación.....	77
Relaciones entre variables.....	79
Coeficiente de correlación.....	80
Regresión lineal simple.....	82
Regresión lineal múltiple.....	87
Sobre las opciones no paramétricas.....	90
Opciones gráficas.....	91
Gráficos básicos.....	91
Gráfico de barras.....	92
Gráficos de área, líneas, pastel y dispersión.....	94
Las barras de error.....	95
Otros gráficos.....	99
Barras apiladas.....	99
Eje X con dos o más variables categóricas.....	100
Gráficos combinados.....	101
Gráfico de doble eje Y.....	102
Gráficos miniatura.....	103
<i>Estadísticas básicas en R.....</i>	<i>105</i>
Nociones sobre el ambiente de R.....	106
Primeros pasos en R.....	106
El formato vectorial.....	111
El formato tabular.....	115
Instalación de paquetes.....	127
Sobre el entorno de RStudio.....	129

Estadística descriptiva en R.....	130
Aplicando estadística descriptiva.....	130
La distribución normal.....	137
Histograma.....	137
Gráfico Q.....	140
Método numérico (curtosis y coeficiente de asimetría).....	141
Método inferencial.....	142
Transformaciones.....	144
Estadística inferencial.....	147
Comparación de proporciones y frecuencias.....	147
Prueba de una proporción.....	148
Pruebas de dos proporciones.....	151
Prueba de bondad de ajuste.....	152
Pruebas de independencia (tablas de contingencia).....	154
Tablas de contingencia 2 x 2.....	154
Tablas de contingencia R x C.....	155
Comparación de medias.....	156
Prueba T para una muestra.....	157
Prueba T para dos muestras independientes.....	159
Prueba T no paramétrica para dos muestras independientes (Wilcoxon o Mann-Whitney).....	161
Prueba T para dos muestras pareadas.....	161
Prueba T no paramétrica para dos muestras pareadas (Wilcoxon).....	163
Análisis de varianza para un factor.....	164
Pruebas de comparaciones múltiples.....	166
Análisis de varianza no paramétrica para un factor (Kruskal-Wallis).....	168
Análisis de varianza para un factor y medidas repetidas.....	168
Análisis de varianza no paramétrico para un factor y medidas repetidas (Prueba de Friedman).....	171
Análisis de varianza para dos factores.....	171
ANDEVA de dos factores con replicación.....	172
ANDEVA de dos factores sin replicación.....	174

Aplicaciones de Estadística Básica

Relaciones entre variables.....	175
Coeficiente de correlación de Pearson.....	175
Correlación no paramétrica (Coeficiente de Correlación de Spearman).....	177
Regresión lineal simple.....	178
Regresión lineal múltiple.....	185
Opciones gráficas.....	190
Gráficos básicos.....	190
Gráfico de barras de una vía.....	190
Gráfico de barras de dos vías.....	203
Gráfico de pastel.....	209
Gráficos de líneas y puntos.....	215
Matriz de gráficos de punto.....	232
Las barras de error.....	233
Otros gráficos.....	238
Gráfico de cajas.....	238
Gráficos de densidad y de violín.....	243
Gráfico de doble eje Y.....	247
Gráfico multipaneles.....	249
Referencias.....	264
Anexos.....	265

Prólogo

La fusión de la estadística con la ciencia de la computación ha sido una mezcla poderosa para facilitar la investigación científica. Los programas de cómputo nos permiten ahorrar tiempo y nos evitan una extenuante labor manual. En la actualidad, para aplicar la estadística, no es necesario memorizar listas inmensas de fórmulas, ni tampoco invertir tiempo haciendo los análisis con lápiz y papel, la tecnología nos ha traído esa plausible bondad a nuestras vidas, en ese sentido, y es necesario aprovecharla.

Esta obra ha sido concebida con la idea de facilitar la aplicación de estadística básica con el uso de los programas Microsoft Excel y R. Por dicha razón, se explican los pasos, funciones y códigos necesarios para la ejecución de cada uno de los procedimientos y pruebas. Se proveen ejemplos enfocados al campo ambiental y se brindan opciones eficientes que permitan la realización de análisis de forma rápida y veraz; aunque también se demanda que el lector ya esté familiarizado con la estadística básica o al menos que tenga buenas referencias bibliográficas para consultar.

Antes de aprender a utilizar Microsoft Excel y R, para implementar aplicaciones estadísticas, estuve buscando manuales o tutoriales, con los cuales pudiera alcanzar este objetivo y me di cuenta de varias cosas: 1. Es difícil encontrar un manual de este tipo, 2. Por lo general los manuales disponibles se encuentran escritos en lengua no española, 3. Muchos autores hacen manuales complejos, pues mezclan varios desempeños (p. ej. combinan procedimientos de administrar datos con los propios análisis), 4. Muchos manuales pasan de lo simple a lo complejo de forma abrupta, lo que le provoca al lector una completa frustración. Dado que no pude encontrar un manual adecuado para mí, he decidido escribir el que yo hubiese querido tener desde el inicio, para aprender a aplicar estadística básica utilizando Microsoft Excel y R.

En mi búsqueda de compartir mis conocimientos con quien lo necesite, he podido concentrar años de entrenamiento, curiosidad y estudio personal en el uso de Microsoft Excel y R para aplicar estadísticas y presentarlas con un lenguaje sencillo. Lo cual también es resultado de una sustancial cantidad de tiempo invertido en aprendizaje mediante prueba y error; búsqueda de solución a momentos de estancamiento; indagaciones sobre análisis menos complicados; consulta a expertos y participación en foros virtuales.

Y aunque esta obra se concentra más en las aplicaciones en los programas, que en explicar los principios estadísticos per se, se pretende inspirar al lector a desarrollar sus capacidades y habilidades en la estadística básica usando programas de cómputo.

Miguel Garmendia Zapata
Profesor Titular
Universidad Nacional Agraria

Introducción

Utilizar programas computacionales para realizar análisis estadístico es una tarea cotidiana en la investigación científica. En el área de las ciencias, el uso de estadística más que fundamental, es mandataria, pues abona a la objetividad, precisión y calidad que todo trabajo científico debe tener; contribuye al procesamiento de información que posteriormente servirá de evidencia para soportar un argumento.

En la investigación, hay mucha diferencia entre llegar a conclusiones mediante la especulación y llegar a conclusiones utilizando estadística. Con esta última, por ejemplo, se pueden aceptar y rechazar hipótesis basadas en criterios meramente numéricos y eso tiene un fuerte impacto en la credibilidad y, al final, en la toma de decisiones. Más en nuestro tiempo, cuando los ordenadores nos permiten aplicar procedimientos estadísticos, que décadas atrás eran tediosos para realizarlos manualmente.

La presente obra reúne años de experiencia, curiosidad, entrenamiento y aplicaciones en investigación de programas estadísticos de uso libre, (como R) o populares como (Microsoft Excel). Esta obra presenta el ABC de las aplicaciones estadísticas en Microsoft Excel (MS Excel) y R de una forma sencilla y accesible para cualquier persona del campo ambiental que tenga interés.

Sin embargo, el lector y quien utilice esta obra, debe tener en mente que necesita conocimientos previos para poder aprovecharla al máximo. En primer lugar, el lector deberá estar familiarizado con el uso básico del programa MS Excel; en segundo lugar, el lector debe tener, a priori, conocimientos básicos de estadísticas. No es objetivo de esta obra enseñar el uso básico del programa MS Excel o las bases teóricas de la estadística; si no, enseñar el uso de los programas MS Excel y R para aplicar estadística básica.

En el caso del programa R, sí se ofrece una introducción muy detallada al programa y se presentan los principios y el ABC de su manejo, debido a que es probable que haya muchos lectores que apenas estén conociendo por primera vez este programa. Por ello, se ofrecen los códigos y los procedimientos de una forma sencilla, ilustrada y muy explicada para el provecho de los entusiastas en el uso de este programa.

Es importante mencionar, que no todas las pruebas estadísticas que se realicen en el programa R se podrán realizar en el programa MS Excel, ya que este último tiene un limitado número de funciones estadísticas en su modo básico (sin instalar complementos adicionales), y, por el contrario, el programa R ofrece una gama increíblemente diversa de funciones en su modo básico (sin instalar paquetes adicionales).

Cuando se habla de estadística básica, tenga en cuenta que lo “básico”, no quiere decir “lo más fácil”, y mucho menos “lo menos importante”; por el contrario, tiene su complejidad, es de suma importancia y es el punto de partida para el aprendizaje de análisis estadísticos más complejos utilizados en el campo de las ciencias ambientales.

Aproveche los conocimientos que se exponen en esta obra como una nueva experiencia en su campo de trabajo, si usted es experto, pues en la vida hay mucho que aprender y de pronto encuentra algo novedoso; si usted no es experto, pero tiene conocimientos básicos de estadística, este libro está hecho para usted; si no es experto y tampoco tiene conocimientos básicos de estadística, también puede utilizar esta obra como punto de partida, pero teniendo a mano alguna literatura de estadística básica a modo de apoyo.

Durante el tiempo que tengo utilizando aplicaciones de MS Excel y R en la estadística, he encontrado otras obras que me han apoyado para incrementar y consolidar mis conocimientos, y ahora han servido de inspiración para este libro. Para MS Excel la obra inspiradora y ahora recomendada es: Carlberg (2011); para R las obras inspiradoras y que recomiendo son: Dalgaard (2002), Teetor (2011) y Zuur et al. (2009).

Selección de la prueba estadística a usar

Para la persona que se está iniciando en el uso de la estadística básica, a veces se le hace complicado seleccionar el tipo de prueba o análisis a utilizar. Esto es algo totalmente normal y comprensible ya que hay muchas pruebas para diferentes situaciones, algunas son pruebas estándares y otras son variaciones a las pruebas estándares. Esto no es algo malo, al contrario, es una enorme fortuna que un investigador tenga una gama de pruebas o análisis para todo tipo de situación, el problema es seleccionar la más apropiada.

Para seleccionar la prueba o análisis más apropiado es recomendable 1. Conocer muy bien la situación, o sea, conocer bien el experimento, ensayo, muestreo, descripción; los objetivos de la investigación que se está desarrollando y en sí, el problema que se quiere resolver y 2. Tener conocimientos sobre estadística básica, en especial en términos de conocer y entender algunos conceptos de importancia como variables, población de estudio, muestra, descripción, normalidad, hipótesis, etc. Con estas dos recomendaciones en mente se hará más fácil encontrar la prueba o análisis más apropiado.

En este capítulo, se ofrece una opción para determinar el tipo de prueba o análisis a realizar, según la información y conocimientos que el lector posea (según recomendaciones arriba). Hay muchas metodologías y criterios para seleccionar pruebas estadísticas, entre ellas los llamados “árboles de decisión” o “claves para selección de pruebas”, los cuales varían de autor a autor. La opción que se ofrece en esta obra está basada mera-

Aplicaciones de Estadística Básica

mente en la experiencia del autor y tratando, en la medida de lo posible, de ofrecer una metodología sencilla, siempre pensando en los novicios de la estadística.

Este capítulo ofrece una clave (basada en criterios estadísticos) para la selección de los análisis y pruebas descritas en esta obra. Adicionalmente, se ofrece un glosario de términos para explicar algunos conceptos y definiciones de una forma sencilla y comprensible (apartando temporalmente la formalidad de términos complicados y fórmulas complejas típicas de la estadística) de tal forma que si el lector no comprende algún término que la clave solicite, pueda recurrir al glosario de términos.

Clave para la Selección de Análisis o Pruebas Estadísticas

Los análisis y pruebas estadísticas que se seleccionarán mediante esta clave, son únicamente aquellos descritos en esta obra y para los cuales se abordan las formas de realizarlos mediante los programas de cómputo Microsoft Excel (2010, 2013, 2016) y R 3.5.1. Se aclara que en esta clave no se abordan todos los análisis o pruebas estadísticas existentes, sino aquellas que el autor ha considerado útiles para el lector a un nivel básico y enfocado en el ámbito ambiental.

La clave está basada en los objetivos que el autor desea lograr, de tal forma que la pregunta inicial y crucial para comenzar la clave es “¿Qué desea hacer?”. A fin de hacer la selección de la repuesta más amigable con el lector, las “opciones de respuestas” enfatizarán responder esa pregunta iniciando con las palabras “deseo”, “quiero”, “quisiera”, etc. El uso de la clave requiere seguir la secuencia de pasos que a continuación se describen.

Paso 1. Hágase la pregunta ¿Qué deseo hacer?

Paso 2. Diríjase a las primeras opciones de respuesta: **1A, 1B, 1C, 1D y 1E**.

Paso 3. Lea cada opción de respuesta (**1A - 1E**) e identifique cuál de ellas describe mejor lo que usted desea hacer con los datos que tiene a mano. Recuerde que si desconoce algún término, búsquelo en el glosario que a continuación se ofrece, o utilice algún recurso bibliográfico extra para aclarar sus dudas.

Paso 4. Si selecciona la opción **1A** la clave le recomendará que realice “Estadística Descriptiva” disponible en MS Excel y en R (notar el punto al lado derecho).

Paso 5. Si selección alguna otra opción, excepto la 1A, la clave lo conducirá a un número que corresponde con otras opciones de respuestas. Por ejemplo, si selecciona la opción

1B, la clave lo conducirá a la respuesta 2, entonces diríjase a las opciones de respuestas que inician con el número 2 o sea **2A, 2B y 2C**.

Paso 6. Proceda con la selección de todas las opciones de respuestas necesarias hasta que llegue a la prueba estadística que se le recomienda utilizar. Adicionalmente, seleccione el programa donde desea aplicar la prueba (EXC para Microsoft Excel o R) y puede dirigirse al índice para explorar en cuál página se explica su implementación.

CLAVE ¿Qué deseo hacer?	EXL	R
1A – Deseo describir los datos en términos de media, moda, error estándar, y otras medidas de posición y dispersión.....Estadística Descriptiva	•	•
1B – Deseo determinar si mis datos cumplen con el supuesto de normalidad.....2		
1C – Deseo comparar datos de frecuencias o proporciones.....4		
1D – Deseo comparar medias de conjuntos de datos de una sola variable.....5		
1E – Deseo comparar relaciones o hacer predicciones entre conjuntos de datos de dos o más variables.....8		
2A – Quisiera explorar la normalidad por métodos gráficos.....Histograma, Gráfico Q	•	•
2B – Quisiera explorar la normalidad por métodos numéricos descriptivos.....Curtosis, Coeficiente de Asimetría	•	•
2C – Quisiera comprobar la normalidad por métodos inferenciales.....3		
3A – Tengo más de 50 observaciones (registros) ($n > 50$).....Prueba de Kolmogorov-Smirnoff		•
3B – Tengo menos de 50 observaciones (registros) ($n \leq 50$).....Prueba de Shapiro-Wilks		•
4A – Deseo determinar si un conjunto de frecuencias se ajustan a un conjunto de proporciones preestablecidas.....Bondad de ajuste (Chi-Cuadrado)	•	•
4B – Deseo comparar una proporción muestreada con una proporción teórica preestablecida.....Prueba de una proporción	•	•

Aplicaciones de Estadística Básica

4C – Deseo comparar dos proporciones muestreadas.....		
.....Prueba de dos proporciones	•	•
4D – Deseo determinar independencia entre las frecuencias de dos variables categóricas con otras dos variables categóricas.....		
.....Tablas de continencia 2 x 2	•	•
4E – Deseo determinar independencia entre las frecuencias de más de dos variables categóricas con más de dos variables categóricas diferentes.....		
.....Tablas de continencia R x C	•	•
5A – Quiero comparar una media con un valor teórico, o dos medias provenientes de dos grupos.....		
6		
5B – Quiero comprar más de dos medias, provenientes de más de dos grupos.....		
7		
6A – Deseo compara una media proveniente de un conjunto de datos con un valor teórico preestablecido.....		
.....Prueba T para una muestra	•	•
6B – Deseo compra dos medias de dos grupos de datos independientes y con distribución normal.....		
.....Prueba T para dos muestras independientes	•	•
6C – Deseo compra dos medias de dos grupos de datos pareados y con distribución normal.....		
.....Prueba T para dos muestras pareadas (apareadas)	•	•
6D – Deseo compra dos medias de dos grupos de datos independientes y que no cumplen el supuesto de distribución normal.....		
.....Prueba de Mann-Whitney		•
6E – Deseo compra dos medias de dos grupos de datos pareados y que no cumplen el supuesto de distribución normal.....		
.....Prueba de Wilcoxon Pareada		•
7A – Deseo compra más de dos medias de varios grupos, un factor y los datos con distribución normal.....		
Análisis de.....		
.....Varianza (ANDEVA) de un Factor	•	•
7B – Deseo compra más de dos medias de varios grupos, un factor y los datos que no cumplen el supuesto de distribución normal.....		
.....Prueba de Kruskal-Wallis		•
7C – Deseo compra más de dos medias de varios grupos, con dos factores.....		
9		
7D – Deseo compara más de dos medias de varios grupos con uno o dos factores y medidas repetidas.....		
.....Prueba de Friedman		•

8A – Quiero explorar la relación entre dos variables con datos que cumplen el supuesto de normalidad.....		
.....Coeficiente de Correlación de Pearson	•	•
8B – Quiero explorar la relación entre dos variables con datos que no cumplen el supuesto de normalidad.....		
.....Coeficiente de Correlación de Spearman		•
8C – Quiero predecir los valores de una variable (dependiente) en función de una o varias variables (independientes).....10	•	•
9A – Quisiera comparar más de dos medias de varios grupos, con dos factores cuyas observaciones están replicadas.....		
.....ANDEVA de dos factores con replicación	•	•
9B – Quisiera comparar más de dos medias de varios grupos, con dos factores cuyas observaciones no están replicadas.....		
.....ANDEVA de dos factores sin replicación	•	•
10A – Quisiera predecir los valores de una variable (dependiente) en función de una variable independiente.....		
.....Regresión Lineal Simple	•	•
10B – Quisiera predecir los valores de una variable (dependiente) en función de diferentes variables independientes.....		
.....Regresión Lineal Múltiple	•	•

Glosario de Términos Relacionados con la Clave

Utilice el siguiente glosario para entender los términos que se muestran en la clave. Las siguientes definiciones están pensadas meramente para el uso de la clave, por eso se emplea el lenguaje más sencillo posible y se presentan ejemplos. Si la información en este glosario no es suficiente para entender los términos, se sugiere recurrir a cualquier libro de estadística básica o bioestadística.

Conjuntos o grupos de datos: Varios valores numéricos que en conjunto conforman “un grupo de datos” con característica, naturaleza y origen común. Ejemplo: En dos parcelas cuadradas establecidas en un bosque, se mide la altura de los árboles (m) que se encuentran dentro de las mismas, en una había seis árboles y en la otra cuatro, de tal forma que los dos conjuntos o grupos de datos por parcela se muestran a continuación:

Parcela 1	Parcela 2
12.3	34.5
24.8	12.8
32.1	18.5
10.6	39.6
18.4	
29.0	

Aplicaciones de Estadística Básica

Datos independientes: Lo contrario de datos pareados. Conjuntos de datos que fueron obtenidos de objetos o lugares diferentes, sin haber ninguna vinculación entre ellos. Ejemplo: Se muestrea el peso de una especie de ave del orden Passeriforme en dos tipos de hábitat Bosque Secundario y Área Agroforestal, se pesaron cuatro pájaros en cada hábitat, los dos conjuntos se presentan a continuación:

Bosque Secundario	Área Agroforestal
12.1	12.5
12.6	13.2
13.4	11.9
12.6	12.2

El conjunto de datos de “Bosque Secundario” es independiente del de “Área Agroforestal” porque ambos provienen de localidades separadas y aves diferentes.

Datos pareados: Lo contrario de datos independientes. Conjuntos de datos que fueron obtenidos de los mismos objetos o lugares, habiendo una vinculación entre ellos. Ejemplo: Se evaluó el daño causado por un insecto a cuatro plantas, contando el número de agallas en las hojas. Se contaron las agallas dos veces en las mismas plantas con cuatro días de separación. Los grupos de datos resultantes se muestran a continuación:

Planta	# Agallas día 1	# Agallas día 4
1	3	4
2	5	8
3	2	4
4	7	9

El conjunto de datos llamado “# Agallas día 1” está vinculado al de “# Agallas día 4” porque ambos fueron tomados a las mismas plantas (objetos). Por ejemplo, para la “planta 1” los valores 3 y 4 representan el número de agallas tomadas en dos momentos diferentes a la misma planta.

Describir: En estadística y por ende en la clave, “describir” se referirá a determinar medias de posición y dispersión en un conjunto de datos. Entre estas se encuentran la media (o promedio), moda, mediana, cuartiles, la desviación estándar, el error estándar, el coeficiente de variación, el número de observaciones, el rango de las observaciones, etc.

Distribución normal: Naturalmente los datos se distribuyen de diferentes formas, la distribución normal es una de ellas y asume una forma de campana (campana de Gauss), en la cual la mayoría de los valores están concentrada en el centro y la minoría en los extremos. Por ejemplo, si tomáramos al azar a 100 personas de un mismo sexo

y edad, y medimos su peso, habrá una considerable cantidad de persona con un peso similar que se agrupa en el centro de la campana, pero habría personas (una minoría) que pesaría muy poco en comparación con la mayoría y que estaría en el lado izquierdo de la campana y personas (otra minoría) que pesarían mucho en comparación con la mayoría y que estarían en el lado derecho de la campana.



Factor: Una cualidad que influye en un objeto u organismo y es medible. Por ejemplo, el crecimiento de las plantas puede estar fuertemente influenciado (no el único, ni el principal) por el factor “luz solar”, de tal forma que a ello se le llama “el factor luz solar”. El crecimiento de un conjunto de plantas se puede medir bajo sombra y a plena luz solar (por ejemplo), de tal forma que al realizar un experimento, ensayo o muestreo se puede afirmar que “el factor luz solar” tiene dos niveles: 1. Bajo sombra (refiriéndose a las plantas que están bajo sombra) y 2. A plena luz solar (refiriéndose a las plantas que están bajo el sol).

Frecuencia: Número de veces que se cuenta un objeto, evento, organismo, número, letra o símbolo. Ejemplo: Numero de huevos en un nido, número de plantas dañadas por un hongo, número de personas de tamaño entre 1.5 y 1.8 metros.

Grupos: Véase conjuntos o grupos de datos.

Independencia (para frecuencias): Se atribuye a variables categóricas que no están asociadas ni guardan relación estadística con otras variables categóricas.

Métodos gráficos: Referido a la elaboración de objetos visuales para representar información (datos), entre ellos histogramas, gráficos barra, pasteles, puntos, líneas, etc.

Métodos inferenciales: Aplicación de pruebas estadísticas para aceptar o rechazar una hipótesis.

Métodos numéricos: Empleo de números para soportar y tomar una decisión.

Aplicaciones de Estadística Básica

Muestreado o muestreada (referido a valores): Valores numéricos o categóricos que se obtienen mediante un muestreo.

Normalidad: Véase distribución normal.

Observaciones (registros): Un dato o un conjunto simultáneo de datos que se le toma a un objeto, organismo, proceso, etc. Ejemplo, se toman dos muestras de suelo y se determina la materia orgánica (MO%), el pH y la humedad (H%). En este caso, cada muestra sería una observación (o registro) y arreglados en una tabla de datos se verían de la siguiente forma:

Muestra	MO%	pH	H%
Observación 1	43.2	5.4	85.4
Observación 2	23.9	6.1	98.2

Una observación correspondería a una fila de datos, así la observación 1 sería conformada por los datos 43.2, 5.4 y 85.4.

Predecir o predicciones: Utilizaremos este término en regresión para estimar los valores de una variable dependiente en función de una variable independiente. Ejemplo: Con la fórmula de regresión $Y=23+0.5(X)$, se puede predecir los valores de la variable Y asignando valores en la variable X; así, cuando $X=2, Y=23+0.5(2)$, o sea $X=2, Y=24$.

Preestablecidas (referido a valores): Valores conocidos o asignados a priori, de forma teórica o sobre la base de estudios anteriores.

Proporciones: Valor de un dato en relación al total del conjunto de datos al que pertenece. En datos de frecuencia una proporción será igual al número de frecuencias de una característica, dividido entre el total de frecuencias de todas las características. Por ejemplo: Se le pregunta a 50 persona si está de acuerdo con un plan de conservación, 42 responden que SÍ y 8 responden que NO, entonces las proporciones de respuestas “SÍ” o “NO” serían:

$$SI = \frac{42}{50} = 0.84$$

$$NO = \frac{8}{50} = 0.16$$

Relaciones (en estadística): Referida a relación entre dos variables, las relaciones pueden ser de tres tipos: 1. Que al incrementar los valores de una variable, también incrementen los valores de la otra variable (relación positiva); 2. Que al incrementar los valores de una variable, decrezcan los valores de la otra variable (relación negativa)

y 3. Que al incrementar los valores de una variable, los valores de la otra variable no manifiesten cambios (sin relación). Ejemplo: Los valores de la temperatura decrecen al aumentar la elevación al nivel del mar, de tal forma que los lugares geográficamente altos son más fríos, esto es un ejemplo de una relación negativa entre las variables temperatura y elevación.

Replicación: Repetición de observaciones en experimentos o muestreos a fin de obtener conjuntos de datos a ser comparados. Por ejemplo: Si se quiere saber la velocidad del crecimiento de una especie de planta sembrada en el suelo, sería incorrecto sembrar solamente una planta y medir el tiempo de crecimiento; lo correcto sería replicar la observación y sembrar, por ejemplo, unas 10 y luego calcular un parámetro que determine el valor de la velocidad de una forma más precisa, puede ser una media, mediana, intervalos de confianza, etc.

Teórico o teórica (referido a valores): Valores numéricos o categóricos (datos) con origen en especulaciones o revisión bibliográfica.

Variable: Una característica medida a un objeto, organismo o proceso. Por ejemplo, el tamaño de un árbol, el potencial de hidrógeno (pH) del suelo, la presencia de mercurio en el agua, el tamaño del fémur de un animal, la velocidad de captación de carbono, etc. todos son ejemplos de variables.

Variables categóricas: Contrario a variables numéricas. Denotan un atributo no medible de forma numérica (de manera general). Por ejemplo: El color de los ojos, comportamiento de un animal, la presencia o ausencia de una plaga, etc.

Variables numéricas: Contrario a variables categóricas. Denotan un atributo medible de forma numérica. Por ejemplo: El número de huevos en nido de aves, el peso en gramos de un animal, la temperatura en grados centígrados, el número de semillas dentro de un fruto, etc.

Variables dependientes e independientes: Dependientes son aquellas variables cuyos valores están en función de otra variable o variables. Por ejemplo, los valores de la temperatura están en función de los valores de la elevación a nivel del mar. Si la elevación varía, la temperatura también, entonces la temperatura es una variable dependiente. Independientes son aquellas variables cuyos valores no están asociados o en función de otra variable o variables. Por ejemplo, los valores de la elevación no están en función de los valores de la temperatura. El que la temperatura varíe no influye en la variación de la elevación, de tal forma que la elevación es una variable independiente, en este caso.

Estadísticas básicas en Microsoft Excel

Microsoft Excel (MS Excel) es un programa de cómputo incluido en el paquete de programas Microsoft Office de Microsoft Corporation, básicamente es una hoja de cálculo con herramientas poderosas para el análisis, manejo y administración de datos. MS Office no es un recurso gratuito, el usuario tiene que activar el producto con periodicidad; sin embargo, el programa MS Excel es gratuito para quien active MS Office.

No es difícil familiarizarse con el ambiente de MS Excel, este sigue la lógica que ha caracterizado todos los programas de MS Office, aunque con ciertas particularidades. Los análisis en MS Excel se pueden hacer de dos formas, uno es con el uso de las funciones que el programa tiene para tales fines y otro es utilizando la hoja de cálculo para realizar análisis con abordaje a mano (o sea estructurar los análisis celda por celda). Evidentemente esta última opción es un poco más complicada y requiere mucho dominio técnico del programa. La mayoría de los análisis en MS Excel los realizaremos con el uso de funciones que el programa trae consigo, y en algún momento, pero con menor frecuencia, utilizaremos algún abordaje a mano con las hojas de cálculos.

Nociones sobre el ambiente de Microsoft Excel

Generalidades sobre una hoja de cálculo

Microsoft Excel es una hoja de cálculo formada por filas y columnas que al superponerse conforman celdas. En cada celda se escribe un valor que puede ser numérico o texto, las operaciones se realizan con la información de cada celda.

Las columnas aparecen nombradas con letras que inician con la A y las filas están nombradas por números que inician con el 1. El nombre (o coordenada) de cada celda se designa con la combinación de la letra de la columna y el número de fila. Por ejemplo, en la figura 1 se ilustra la sección de una hoja de cálculo en la cual el número 25, que en ella aparece, se encuentra ubicado en la celda B5 (Columna B, Fila 5) y la palabra “Cuenta” se encuentra en la celda C6 (Columna C, Fila 6). Observemos en la misma figura que la celda A1 es la “celda activa”, esto se determina por el recuadro grueso que la rodea, la celda activa es donde se insertarán los datos o se aplicarán las funciones. Para activar una celda solo hacemos clic sobre la misma o la seleccionamos con las flechas direccionales.

Podemos activar una o varias celdas de forma continua haciendo clic en una y arrastrando el cursor hacia las celdas que se desean seleccionar. Cuando seleccionamos varias celdas de forma continua vertical (varias celdas en columna) u horizontalmente (varias celdas en fila) se selecciona un “rango de datos”, los rangos de datos se denotan con el nombre de la celda que inicia el rango y el nombre de la que lo finaliza, separadas por dos puntos (:). Notemos el rango de datos en la columna D (Figura 1), el cual se representaría por D1:D4.

También podemos seleccionar simultáneamente varias celdas individuales o en conjuntos en diferentes partes de la hoja de cálculo, manteniendo presionada la tecla Control (Ctrl). Por ejemplo, podemos seleccionar la celda A1, luego presionar Ctrl y seleccionamos el rango D1:D4 a como se observa en la figura 1.

		Columna A	Columna B	Columna C	Columna D
		A	B	C	D
Fila 1	1				
Fila 2	2				
Fila 3	3				
Fila 4	4				
Fila 5	5		25		
Fila 6	6			Cuenta	

Figura 1. Ilustración de una sección de cuatro columnas y seis filas de una hoja de cálculo de Microsoft Excel.

Un archivo de MS Excel se denomina “Libro” y las pestañas que aparecen en la parte inferior izquierda se denominan “Hojas”. Al hacer clic sobre el signo más (+) dentro de un círculo a la par de las hojas, anexamos una hoja nueva con el nombre de “Hoja 2” para diferenciarla de la que aparece por defecto (Hoja 1) (Figura 2). Es posible cambiar el nombre de la hoja haciendo doble clic sobre el nombre que tiene por defecto.

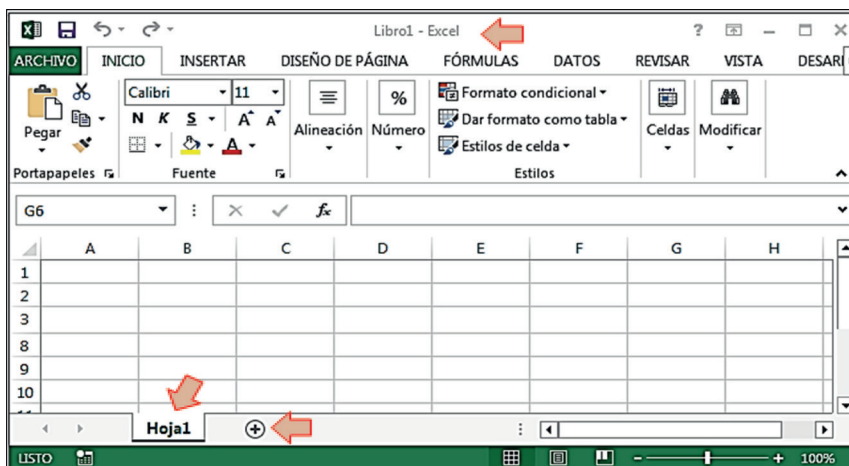


Figura 2. Ejemplo de un libro de MS Excel con todos sus componentes. La flecha en la parte superior está indicando el nombre del libro, el cual es el mismo nombre del archivo. Las dos flechas de abajo están indicando el nombre de la hoja y el ícono para insertar una nueva hoja de cálculo.

Aplicaciones de Estadística Básica

Operaciones y funciones

Microsoft Excel puede servir como una simple calculadora, aunque sus aplicaciones van más allá de esta forma de usarlo, pero de forma básica se pueden hacer operaciones como suma (+), resta (-), multiplicación (*) y división (/). En la figura 3 se ilustra el uso de las funciones para dichas operaciones, en el ejemplo se tienen dos números (el 3 y el 5) para aplicar las cuatro operaciones. Estas se aplicarán a modo de fórmula, y en MS Excel una fórmula se inicia escribiendo el símbolo igual (=) en la celda donde se desea que aparezca el resultado.

La fórmula no incluye el número directamente, sino la coordenada de la celda donde están los números que se utilizarán para el cálculo, en el primer caso el número 3 está en la celda A1 y el número 5 en la celda B1, de tal forma que para sumarlos debemos utilizar la fórmula $=A1+B1$. Al presionar Enter, el programa realiza la sumatoria y coloca el resultado en la celda C1, que fue donde se escribió la fórmula (Figura 3). De igual forma procedemos para las otras operaciones y, en el caso del ejemplo, solo cambiaremos el número de las filas y las funciones matemáticas.

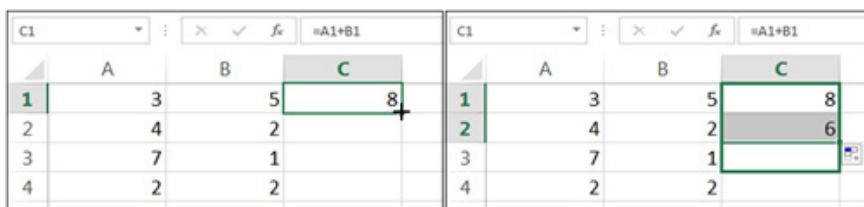
Las funciones aplican no solo a lo largo de las filas, sino a lo largo de las columnas, por ejemplo si se quieren sumar los números 3 de la columna A, basta con escribir la fórmula $=A1+A2+A3+A4$ en la celda donde obtenemos el resultado, si en lugar de sumar deseamos multiplicar, solamente sustituimos el símbolo de suma (+) por el de multiplicación (*), resultando la fórmula $=A1*A2*A3*A4$, y así sucesivamente con las funciones de resta (-) y división (/). Para el caso particular de la suma, MS Excel nos ofrece la función "SUMA()" a la cual solamente asignamos el rango de celdas donde están los números y nos resulta la fórmula $"=SUMA(A1:A4)"$, con la cual el programa realiza la sumatoria (de una forma más automatizada) de todos los valores dentro de dicho rango (Figura 3).

	A	B	C	D
1	3	5	$=A1+B1$	
2	3	5	$=A2-B2$	
3	3	5	$=A3*B3$	
4	3	5	$=A4/B4$	
5	$=A1+A2+A3+A4$			
6	$=SUMA(A1:A4)$			

	A	B	C	D
1	3	5	8	
2	3	5	-2	
3	3	5	15	
4	3	5	0.6	
5	12			
6	12			

Figura 3. Ilustración de las operaciones matemáticas de suma (+), resta (-), multiplicación (*) y división (/) en una hoja de cálculo de MS Excel. Adicionalmente se ejemplifica el uso de la función "SUMA()". A la derecha se presentan los resultados de los cálculos.

Si quisiéramos ejecutar una misma operación fila por fila o columna por columna para un conjunto de datos, MS Excel nos ofrece una herramienta de relleno. En la esquina inferior derecha de cada celda activa aparece un cuadro que al ponérsele el cursor encima se forma un signo más (+), allí hacemos clic derecho, mantenemos presionado y arrastramos la selección hasta el último número donde se quiere que se ejecute la operación. En la figura 4 se ilustra un ejemplo donde se ejecutó la operación de suma de los números 3 y 5 mediante la fórmula =A1+B1, y luego se utilizó la herramienta de relleno con la que el programa copia la operación y la ejecuta a cada fila de datos subsiguiente (4 y 2, 7 y 1, 2 y 2). Notemos que es necesario hacer la operación una vez en la primera celda para que el programa reconozca cuál es la operación y la aplique para los otros casos.



	A	B	C
1	3	5	8
2	4	2	6
3	7	1	
4	2	2	

Figura 4. Ilustración del uso de la herramienta de relleno para automatizar una operación en varias filas. A la izquierda se observan dos conjuntos de datos, uno en la columna A y otro en la columna B y se suman los números por fila. Notemos la fórmula en el recuadro superior donde se escriben las funciones (fx). A la derecha se representa el arrastre de la operación, dando como resultado el número 6 para la fila 2 (4 + 2) y así sucesivamente se suman las restantes filas.

Podemos encontrar más funciones en la secuencia de opciones Inicio>Autosuma>Más funciones...>Seleccionar una función. En la opción llamada “O seleccionar una categoría:” se pueden seleccionar las funciones usadas recientemente, todas las funciones o funciones específicas por categorías: financieras, matemáticas y trigonometrías, estadísticas, lógica, etc. (Figura 5).

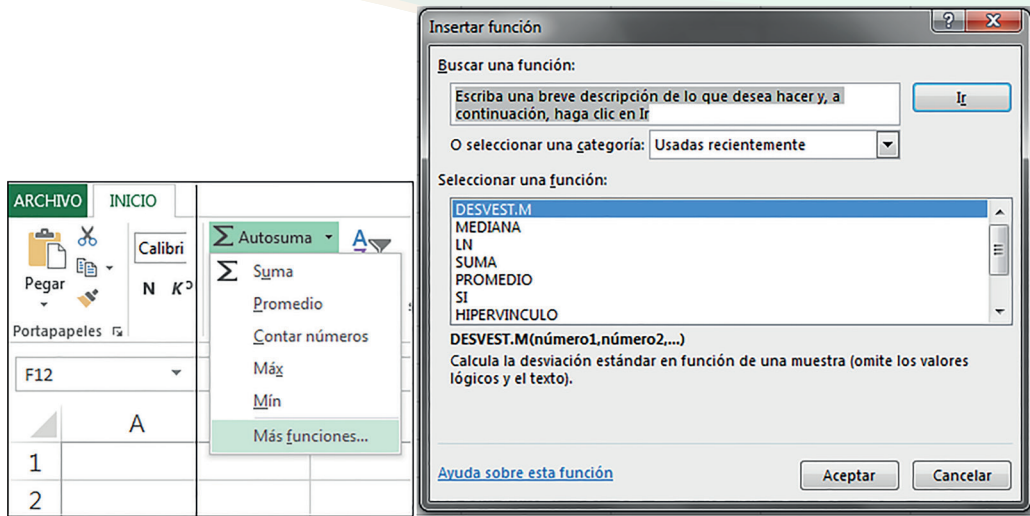


Figura 5. Secuencia de opciones para buscar más funciones, las que se incluyen dentro de las categorías financieras, matemáticas y trigonometría, estadística, lógica, etc.

Fórmulas en MS Excel

Es común en MS Excel el uso de fórmulas para ejecutar funciones que no están en el programa. En principio, es necesario conocer la fórmula y transformarla a una expresión que MS Excel pueda interpretar. Esto incluye el uso de los nombres de las celdas donde están los valores y de varios símbolos. En esto se debe de tener sumo cuidado, cualquier error de escritura puede resultar en un grave error de cálculo, siempre se recomienda validar las fórmulas con el uso de otros medios (calculadoras, programas, etc.) al menos la primera vez que se usan. A continuación se presentarán varios ejemplos de cómo transformar una fórmula convencional a una expresión legible por MS Excel.

Si quisiéramos determinar el área de un círculo de 25 m de radio, podemos reescribir la fórmula de área de círculo en una forma legible por MS Excel. La fórmula a aplicar sería expresada como: $A = \pi r^2$, en MS Excel dicha fórmula la deberíamos expresar como $=PI()*25^2 = 1963.495408$ o $3.1416*25^2 = 1963.5$. La función "PI()" (con los paréntesis vacíos) le indica al programa el uso del valor de π (3.1416...); el símbolo "*" indica una multiplicación y el símbolo "^" eleva el número que está a su izquierda (25) a la potencia deseada y escrita a su derecha (2).

Otros ejemplos:

Se pretende calcular el porcentaje de germinación de un grupo de 45 semillas, de las cuales 21 germinaron.

Fórmula: $\% = \frac{21}{45} \times 100$

Fórmula en MS Excel: $= (21/45)*100$

Si el número 21 se encontrase en la celda B2 y el número 45 se encontrase en la celda B3, la fórmula en MS Excel se expresaría como: $= (B2/B3)*100$

Notar el símbolo igual para abrir la fórmula y los paréntesis para separar las operaciones.

Para calcular el error estándar de una muestra se utiliza la fórmula: $EE = \frac{\sigma}{\sqrt{n}}$;

Donde,

EE= Error estándar

σ = Desviación estándar

n= número de muestras

Si $\sigma = 5.6$ y $n = 6$, la fórmula en MS Excel se expresaría como: $=5.6/RAIZ(6)$, lo que es igual a 2.2861.

Si el 5.6 se encuentra en la celda D3 y el 6 se encuentra en la celda G3, la fórmula en MS Excel se escribiría de la siguiente manera: $=D3/RAIZ(G3)$.

Notemos el uso de la función “RAIZ()” para calcular la raíz cuadrada del número, si desea conocer más sobre las funciones en MS Excel, diríjase a la opción de “Ayuda” y escriba en el buscador la palabra “Funciones”, MS Excel le mostrará todas las funciones por categoría. También se puede dirigir a la opción de “Insertar función” descrito en la figura 5.

La fórmula para calcular el estadístico Z para comparar dos proporciones es algo más compleja que las anteriores, expresarla en una fórmula de MS Excel requiere mucha cautela. La fórmula de Z se expresa a continuación:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p} \bar{q}}{n_1} + \frac{\bar{p} \bar{q}}{n_2}}}$$

Aplicaciones de Estadística Básica

Donde,

\hat{p}_1 = Proporción 1 = 0.44 = ubicado hipotéticamente en la celda B22

\hat{p}_2 = Proporción 2 = 0.82 = ubicado en la celda B23

\bar{p} = Proporción combinada = 0.62 = ubicado en la celda B24

\bar{q} = 1-Proporción combinada = 0.38 = ubicado en la celda B25

n_1 = Cuenta para la variable 1 = 77 = ubicado en la celda B18

n_2 = Cuenta para la variable 2 = 68 = ubicado en la celda C18

En primera instancia, sustituiremos, a modo de ejemplo, los símbolos de la fórmula por los nombres de las celdas donde se encuentran los valores correspondientes a cada símbolo:

$$Z = \frac{B22 - B23}{\sqrt{\frac{B24 \times B25}{B18} + \frac{B24 \times B25}{C18}}}$$

Sin embargo, esta fórmula no puede ser utilizada en MS Excel, así que la transformamos a una expresión legible por MS Excel:

$$=(B22-B23)/RAIZ((B24*B25/B18)+(B24*B25/C18))$$

Si sustituimos los valores correspondientes a cada nombre de celda, el resultado sería: -4.73. Notemos el uso de los paréntesis para separar las operaciones dentro de la formula, no es lo mismo expresar “B24*B25/B18+B24*B25/C18” que “(B24*B25/B18)+(B24*B25/C18)”, en la primera puede que el programa este efectuando las operaciones en cadena, o sea, que primero multiplique B24 por B25, luego ese resultado lo divida entre B18 y el siguiente resultado lo sume con B24 hasta terminar la cadena de operaciones, eso daría un resultado erróneo.

En el segundo caso, le estamos diciendo al programa que solamente después de realizar los dos conjuntos de operaciones (que están entre paréntesis) se continúa con la suma de los productos de ambas. Si sustituimos los nombres de las celdas por números, podemos dilucidar las diferencias en resultados:

$$0.62*0.38/77+0.62*0.38/68 = 0.0035$$
$$(0.62*0.38/77)+(0.62*0.38/68) = 0.0065$$

A veces MS Excel puede utilizar la lógica para realizar las operaciones y dar resultados correctos sin incluir los paréntesis, pero no se garantiza que eso suceda correctamente en todos los casos. Por lo que es mejor asegurarse un resultado correcto, al separar las operaciones individuales utilizando paréntesis.

La herramienta de análisis de datos

MS Excel tiene una herramienta donde concentra todos los análisis estadísticos incluidos en el programa (tantos de estadística descriptiva como de inferencia). La opción no se encuentra disponible cuando el programa se instala, hay que hacerla disponible mediante los pasos, que a continuación se describen.

Diríjase a la opción “Archivo” del programa MS Excel (Paso 1) (Figura 6), haga clic en “Opciones” (Paso 2), seleccione la opción “Complementos” (Paso 3), verifique que donde dice “Administrador:” esté seleccionada la opción “Complementos de Excel” (Paso 4) y haga clic en “Ir...” (Paso 5), aparecerá un cuadro de diálogos donde tendrá que poner el check en la opción “Herramientas para análisis” (Paso 6), más aceptar (Paso 7).

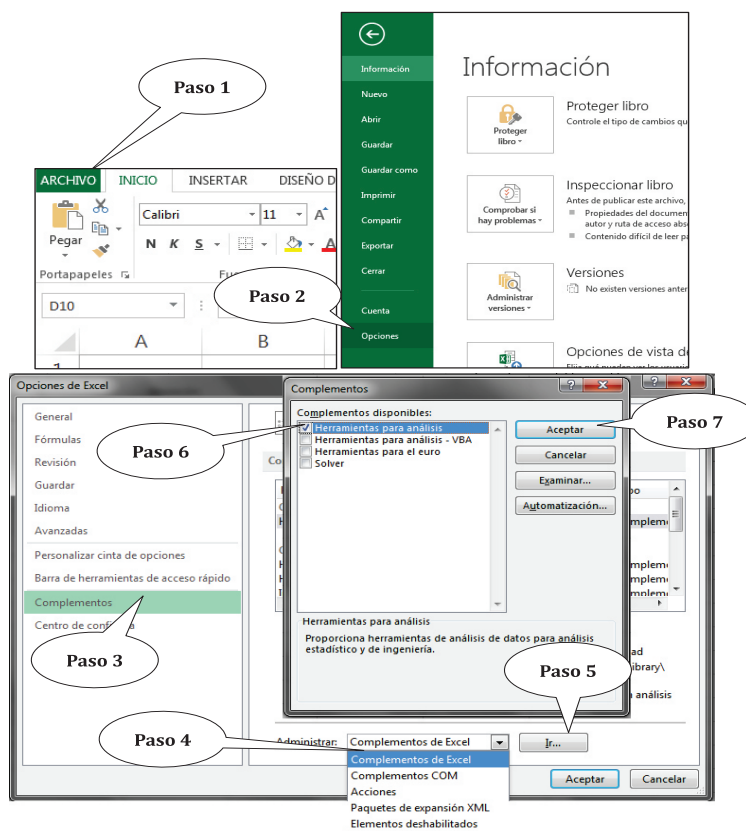


Figura 6. Ilustración del procedimiento para hacer disponible la opción “Herramientas para análisis” en MS Excel, con la cual se realizarán los análisis estadísticos.

Seguidamente diríjase a la pestaña llamada “Datos” y verifique que aparece un ícono nuevo con la descripción “Análisis de datos” (Figura 7).

Aplicaciones de Estadística Básica



Figura 7. Visualización de la herramienta de “Análisis de Datos”, dentro de las opciones de “Datos”, después de hacerla disponible.

Complementos para MS Excel

Es necesario mencionar que las funciones básicas de estadísticas y de análisis en MS Excel son limitadas, por lo cual se han creado complementos para fortalecer esas funciones. Por ejemplo, el complemento “Real Statistics Using Excel” es gratuito, se puede instalar, hacer disponible y realizar análisis que no serían posibles con la extensión “Herramientas para análisis”, que ofrece MS Excel por defecto o que por otra parte sería tedioso hacerlo manualmente en las hojas de cálculo. El complemento se encuentra en el sitio Web <http://www.real-statistics.com/>

Otros complementos involucran una suscripción y un pago de licencia, los que también nos ofrecen análisis y opciones gráficas, que no están disponibles entre las funciones básicas de MS Excel. Entre ellos, el paquete estadístico llamado “Analyse-it” (visitar: <https://analyse-it.com/>) y el llamado “XLSTAT” (visitar: <https://www.xlstat.com/es/>). Otro complemento es “MegaStat Software”, de marca registrada por J.B.Orris, Butler University, el cual es parte de un libro llamado “Essentials of Business Statistics” (Estadísticas Esenciales para Negocios), su uso tiene restricciones, está disponible en: http://highered.mheducation.com/sites/0071339604/student_view0/megastat_software.html. Más complementos gratis o pagados se pueden explorar en el sitio Web de Microsoft Excel Add-ins: <https://www.add-ins.com/index.htm>.

Para efecto de esta guía, solo se utilizará la opción “Herramientas para análisis” (Análisis de Datos) que MS Excel trae consigo, pero se debe hacer disponible. Otros complementos no serán abordados y su uso queda a criterio del lector.

Estadística descriptiva en Microsoft Excel

Aplicando estadística descriptiva

De manera general se podría decir que la estadística descriptiva, como su nombre lo indica, describe los datos en términos de su tendencia central, dispersión y forma. Esto es

imprescindible para conocer la naturaleza de los datos y aplicar subsiguientes pruebas estadísticas. Para ejemplificar el uso de las estadísticas descriptivas, estructuraremos un pequeño conjunto de datos correspondientes a información de pH (potencial de hidrógeno) del suelo en cinco puntos, en un área agrícola (Cuadro 1). Se ofrece un conjunto de datos lo suficientemente pequeño, para que el lector pueda escribirlos en una hoja de cálculo de MS Excel y practicar el procedimiento.

Cuadro 1. Datos para ejemplificar el cálculo de estadísticas descriptivas en MS Excel.

No PUNTO	pH
1	4.7
2	5.3
3	5.9
4	4.9
5	4.6

Para aplicar estadística descriptiva al conjunto de datos mostrado anteriormente, nos dirigimos a la pestaña “Datos” (Paso 1) (Figura 8), en el extremo derecho de las opciones desplegadas encontraremos la herramienta “Análisis de datos” (Paso 2) (recorde-mos que primero tenemos que hacer disponible esta herramienta), aparecerá un cuadro de diálogo donde tendremos que dirigirnos a la opción “Estadística descriptiva” (Paso 3). Al finalizar los pasos anteriores nos aparecerá un cuadro de diálogo, el cual solicita instrucciones para el análisis.

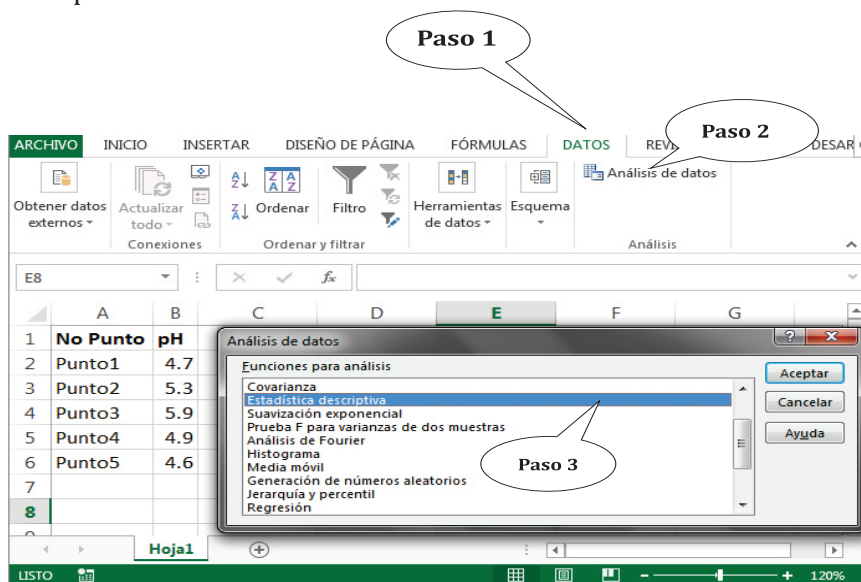


Figura 8. Ilustración de los pasos para aplicar la opción de “Estadística descriptiva”.

Aplicaciones de Estadística Básica

Hacemos clic en la caja (flecha roja) que está al lado derecho donde dice “Rango de entrada” (Paso 4) (Figura 9), el programa minimiza dicha ventana y da lugar a que se seleccione el rango de datos (incluyendo el nombre de la columna) (Paso 5). Hacemos clic nuevamente en la flecha roja de la caja minimizada (Paso 6).

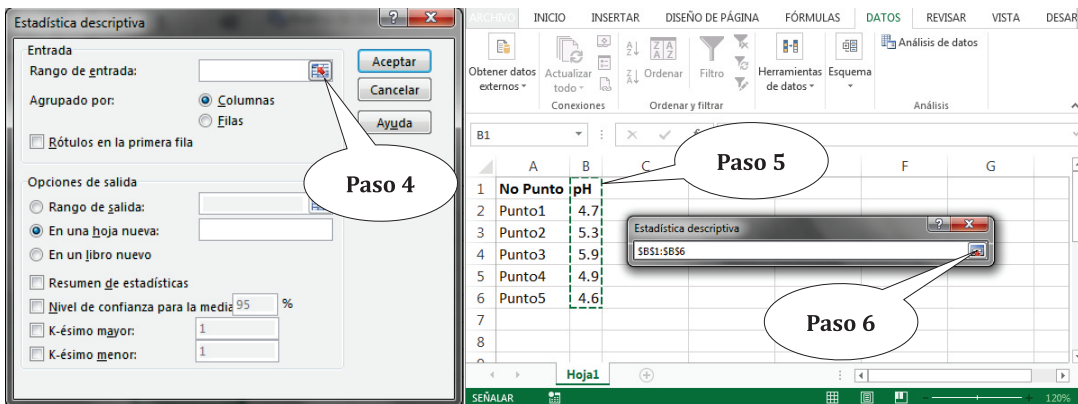


Figura 9. Ilustración de la selección de la información a utilizarse en el análisis.

En la figura 10 se observa el rango de entrada seleccionado, lo observamos en la expresión “\$B\$1:\$B\$6” que en el lenguaje de MS Excel significa que están seleccionados los datos, desde la celda B1 (\$B\$1) a la celda B6 (\$B\$6).

Siguiendo con los pasos, ponemos check en “Rótulos en la primera fila” (Paso 7) lo que le indica al programa que la primera celda del rango es el nombre de la columna (pH), y que por lo tanto lo excluya de los análisis. Luego hacemos clic en la opción “Rango de salida” (Paso 8) donde le especificaremos al programa en qué parte de nuestra hoja de cálculo colocará la tabla de resultados. Si decidimos que los resultados se presenten en una hoja nueva, se puede dejar la opción por defecto (En una hoja nueva). Seguidamente hacemos clic en la caja con la flecha roja (Paso 9) que está al lado derecho, lo que nos permitirá definir la celda, que será la esquina superior izquierda de la tabla de salida de los resultados (Paso 10) y regresamos al cuadro de diálogo haciendo clic de nuevo en la flecha roja (Paso 11).

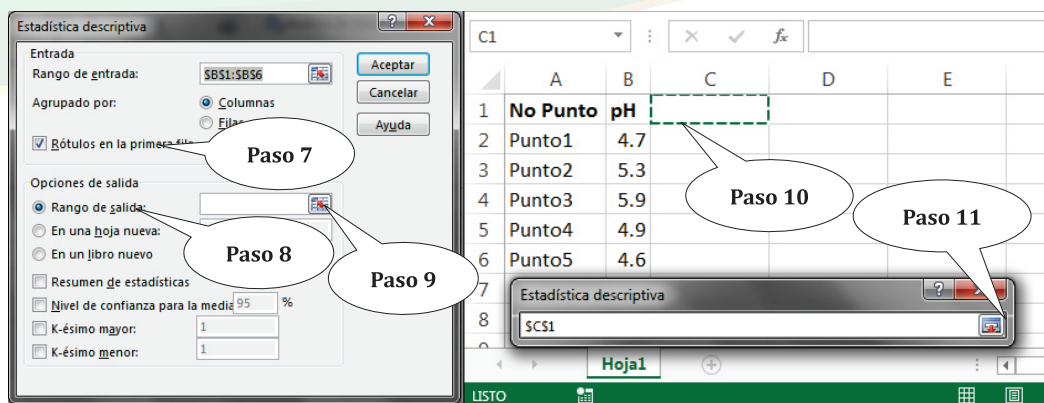


Figura 10. Continuación de la ilustración de la secuencia de pasos para aplicar estadística descriptiva.

Observemos en la figura 11 en “Rango de salida”, que el programa ya tiene la coordenada de donde se desplegará la tabla (\$C\$1 correspondiendo a la celda C1). Como pasos finales ponemos check en “Resumen de estadísticas” (Paso 12) y en “Nivel de confianza para la media” (Paso 13), y lo dejamos en su valor por defecto (95%), finalmente seleccionamos “Aceptar”.

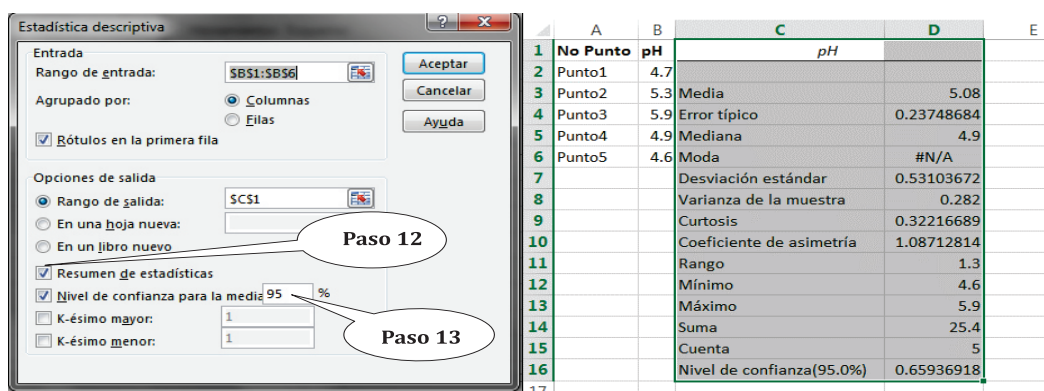


Figura 11. A la izquierda se muestran los dos últimos pasos para ejecutar la operación. A la derecha se muestra el cuadro de resultados.

La figura 11 presenta el cuadro de resultados de la operación “Estadística descriptiva”. A como se observa, inicia con un título, correspondiente al nombre de la columna de datos, luego presenta dos columnas, en la primera se muestran los nombres de los esta-

Aplicaciones de Estadística Básica

dísticos descriptivos y en la segunda sus valores. Con esta opción, el programa calculó medidas de tendencia central como la media, la mediana y la moda; medidas de dispersión como el error típico (error estándar), la desviación estándar, la varianza; medidas de forma como curtosis y coeficiente de asimetría; además de otros datos descriptivos como el rango, el valor mínimo y máximo, la suma y cuenta de los valores, y el nivel de confianza (95.0%). El investigador selecciona aquellas medidas que les sean más útiles para describir sus datos. Para interpretar los valores se sugiere la lectura de cualquier libro de estadística.

Sin embargo, haremos hincapié en el nivel de confianza, la curtosis y el coeficiente de asimetría, ya que la interpretación de estos requiere algunos pasos extras. Con el valor obtenido por el nivel de confianza (calculado para una distribución t) podemos obtener los intervalos de confianza para un $\alpha = 0.05$. Estos los construimos sumando y restando dicho valor a la media (media \pm el nivel de confianza), o sea 5.08 ± 0.66 , lo que daría como resultado:

$$5.08 - 0.66 = 4.42$$

$$5.08 + 0.66 = 5.74$$

Por lo que se interpreta que tenemos un 95% de confianza de que la media del pH se encuentre 4.42 y 5.74.

La curtosis la debemos de interpretar comparando el valor obtenido con unos rangos de valores que están asociado con el tipo de curtosis. En sí, la curtosis mide el grado de agudeza de la distribución de los datos y hay de tres tipos: Leptocúrtica, Mesocúrtica y Platicúrtica. Al igual que la curtosis, el coeficiente de asimetría se interpreta comparado el valor obtenido con unos rangos de valores que están asociado con el tipo de asimetría, esta permite determinar la forma cómo se distribuyen los datos y hay tres tipos: Asimétrica Negativa, Simétrica y Asimétrica Positiva (Figura 12).

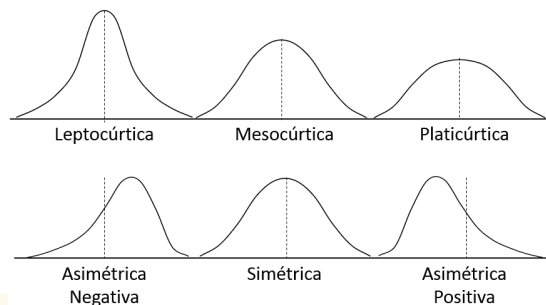


Figura 12. Formas que adquiere la distribución de datos de manera general. Arriba las tres formas de curtosis y abajo las tres formas de asimetría. La línea vertical punteada muestra la posición relativa de la media.

El cuadro 2 muestra los rangos de valores con los que tenemos que comparar los resultados de la curtosis y el coeficiente de asimetría para determinar los tipos.

Cuadro 2. Rango de valores (criterio) para definir los tipos de curtosis o asimetría. La curtosis es basada en cuartiles y percentiles.

CURTOSIS (k)		ASIMETRÍA (a)	
Criterio	Tipo de curtosis	Criterio	Tipo de simetría
$k < 0.263$	Platicúrtica	$a < 0$	Asimétrica negativa
$k = 0.263$	Mesocúrtica	$a = 0$	Simétrica
$k > 0.263$	Leptocúrtica	$a > 0$	Asimétrica positiva

Hasta acá, hemos visto la aplicación de la opción de estadística descriptiva a una variable (en una columna de datos), si se tiene más de una variable, simplemente procedemos con los mismos pasos ya vistos y solo ampliamos el rango de selección de los datos, en el paso correspondiente.

Si es de nuestro interés determinar las medidas descriptivas de manera individual, entonces se pueden utilizar las funciones estadísticas de MS Excel. Por ejemplo, para calcular la media (promedio) de los datos en el cuadro 1, seleccionamos la celda donde se va presentar el resultado del cálculo, escribimos el signo igual “=” y el nombre de la función, en este caso “PROMEDIO()” (Figura 13). Notemos que el programa nos presenta todas las funciones que coinciden con la palabra escrita, de tal forma que con doble clic seleccionamos la indicada.

Al seleccionar la función, el programa pone en la celda de resultados la expresión “=PROMEDIO()” y muestra una pequeña etiqueta, indicando que es necesario ingresar el rango de números (número 1...), para lo cual se seleccionan todos los números y se presiona la tecla Enter.

	A	B	C		A	B	C		A	B	C
1	No Punto	pH		1	No Punto	pH		1	No Punto	pH	
2	Punto1	4.7		2	Punto1	4.7		2	Punto1	4.7	
3	Punto2	5.3		3	Punto2	5.3		3	Punto2	5.3	
4	Punto3	5.9		4	Punto3	5.9		4	Punto3	5.9	
5	Punto4	4.9		5	Punto4	4.9		5	Punto4	4.9	
6	Punto5	4.6		6	Punto5	4.6		6	Punto5	4.6	
7		=PROMEDI		7		=PROMEDIO(7		=PROMEDIO(B2:B6	
8		PROMEDIO		8		PROMEDIO(número1, [número2], ...)		8		PROMEDIO(número1, [número2], ...)	
9		PROMEDIO.SI									
10		PROMEDIO.SI.CONJUNTO									
		PROMEDIOA									

Figura 13. Ilustración del proceso para calcular la media (promedio) utilizando la función “PROMEDIO()” en MS Excel.

Aplicaciones de Estadística Básica

Así como calculamos la media con el uso de una función específica, también podemos calcular las otras medidas descriptivas mediante sus funciones, las cuales se muestran en el cuadro 3.

Cuadro 3. Lista de funciones para aplicar estadística descriptiva en MS Excel. También se ofrece una descripción del uso de cada función. Se utiliza como ejemplo el rango de datos B2:B6 (Figura 13).

MEDIDAS	FUNCIÓN	DESCRIPCIÓN
Media	=PROMEDIO(B2:B6)	Calcula el promedio para el rango de datos B2 – B6.
Error típico	=0.53/RAIZ(5)	Calcula el error estándar. No hay función directa, se calcula con la fórmula: desviación estándar dividida por la raíz cuadrada del número de observaciones.
	=B11/RAIZ(B19)	Otra opción: Cuando la desviación estándar (B11) y el número de observaciones (B19) tienen celdas definidas en la hoja de cálculo.
	=MEDIANA(B2:B6)	Calcula la mediana para el rango de datos B2 – B6.
Moda	=MODA.UNO(B2:B6)	Calcula la moda para el rango de datos B2 – B6.
Desviación estándar	=DESVEST.M(B2:B6)	Calcula la desviación estándar de la muestra para el rango de datos B2 – B6.
Varianza	=VAR.S(B2:B6)	Calcula la varianza de la muestra para el rango de datos B2 – B6.
Curtosis	=CURTOSIS(B2:B6)	Calcula la curtosis para el rango de datos B2 – B6.
Coefficiente de asimetría	=COEFICIENTE.ASIMETRIA(B2:B6)	Calcula el coeficiente de asimetría para el rango de datos B2 – B6.
Rango	=MAX(B2:B6) - MIN(B2:B6)	Diferencia entre el valor máximo y mínimo de un conjunto de datos.
Máximo	=MAX(B2:B6)	Selecciona el valor máximo de un conjunto de valores.
Mínimo	=MIN(B2:B6)	Selecciona el valor mínimo de un conjunto de valores.
Suma	=SUMA(B2:B6)	Suma los valores del rango de datos B2 – B6.
Cuenta	=CONTAR(B2:B6)	Cuenta los valores del rango de datos B2 – B6.
Nivel de confianza (95%)	=INTERVALO.CONFIANZA.T(0.05, 0.53, 5)	Calcula el nivel de confianza. Se tiene que especificar el nivel de significancia (α) que por lo general es 0.05, la desviación estándar (0.53) y el número de observaciones (5).
	=INTERVALO.CONFIANZA.T(0.05, B11, B19)	Otra opción: Cuando la desviación estándar (B11) y el número de observaciones (B19) tienen celdas definidas en la hoja de cálculo.

Las funciones estadísticas también se encuentran en MS Excel en la pestaña “INICIO” dentro de las funciones denotadas con el símbolo “ Σ Autosuma”, desplegando la flecha de opciones.

La distribución normal

Evaluación de la distribución normal

Un conjunto de datos se puede distribuir de diversas formas, una de ellas es la distribución normal que adquiere una forma simétrica como la de una campana, donde hay valores muy frecuentes en el centro y valores muy raros en los extremos, como la mostrada en la figura 12, denominadas mesocúrtica y simétrica. Es necesario conocer si la distribución de los datos es normal, pues las pruebas paramétricas tienen como pre-misa esta característica; si los datos no poseen dicha característica, podemos seguir dos

vías: 1. Podemos transformarlos o 2. Podemos utilizar pruebas no paramétricas. A continuación se mostrarán tres formas de determinar si un conjunto de datos se ajusta a la distribución normal, entre los cuales están dos formas gráficas y dos formas numéricas.

En la figura 14 se presentan los datos de mediciones de altura (en cm) a 10 personas, con ellos evaluaremos el ajuste de los datos a una distribución normal. Emplearemos tres formas sencillas de hacer dicha evaluación, gráficamente mediante un histograma y un gráfico Q, y numéricamente mediante la exploración de la forma de la curva, con las funciones curtosis y coeficiente de asimetría.

	A	B		A	B	C	
1	Personas	Altura		1	Personas	Altura	Clase
2	Persona1	180		2	Persona1	180	150
3	Persona2	172		3	Persona2	172	160
4	Persona3	160		4	Persona3	160	170
5	Persona4	173		5	Persona4	173	180
6	Persona5	165		6	Persona5	165	
7	Persona6	168		7	Persona6	168	
8	Persona7	143		8	Persona7	143	
9	Persona8	162		9	Persona8	162	
10	Persona9	177		10	Persona9	177	
11	Persona10	153		11	Persona10	153	

Figura 14. Registros de alturas (en centímetro) de 10 personas en una hoja de cálculo de MS Excel. A la derecha se observa el mismo conjunto de datos, pero con la adición de una variable nueva llamada “Clase” para crear un histograma.

Histograma

Un histograma es básicamente un gráfico de barra, en el cual se muestra en el eje X una variable numérica donde se distribuyen de forma ascendente los números o los rangos de números y en el eje Y se muestran las frecuencias de ocurrencias de esos números o rangos de números.

La variable “Clase”, en la figura 15, será utilizada para dividir el conjunto de los datos que van a ser presentados en el histograma, a esto también se le conoce como rangos de datos. La forma de estructurar los rangos depende básicamente del investigador, pero se debe cuidar que los rangos incluyan a los valores extremos (mínimo y máximo).

En el ejemplo, el primer valor de la clase es 150, representa al rango $\infty \leq 150$, o sea que contará todos los valores menores e igual a 150 y esa cuenta será la frecuencia expresada en el gráfico; así, con el número 160 en la clase, el programa contará la frecuencia de

Aplicaciones de Estadística Básica

todos los valores entre 151 y 160 (incluyendo al 160) y así sucesivamente se continúa con los otros valores de la variable “Clase”.

Crearemos el histograma utilizando la herramienta de “Análisis de datos” que se encuentra en la pestaña Datos>Análisis de datos, luego seleccionaremos la opción “Histograma” (Paso 1) (Figura 15), se desplegará un cuadro de diálogos solicitando información para generar el histograma. En “Rango de entrada” hacemos clic en la flecha roja de la caja correspondiente, el cuadro de diálogo se minimizará y nos dará lugar a que se seleccionen los datos, únicamente seleccionaremos los valores numéricos en la variable “Altura”, junto con el encabezado de la columna (Paso 2), luego volvemos a hacer clic en la flecha roja para maximizar el cuadro de diálogo y donde dice “Rango de clase” procederemos a seleccionar los datos y el encabezado de la variable “Clase” (Paso 3). Recordemos poner check en “Rótulos” (Paso 4) y seleccionar la opción de salida, en el caso de este ejemplo seleccionaremos la opción “Rango de salida” (Paso 5), pero el lector es libre de explorar las otras opciones; seguidamente definiremos el rango de salida haciendo clic en la flecha roja y seleccionando la celda donde queremos que se presente la tabla de resultados (Paso 6), finalmente ponemos check donde dice “Crear gráfico” (Paso 7) y hacemos clic en “Aceptar” (Paso 8).

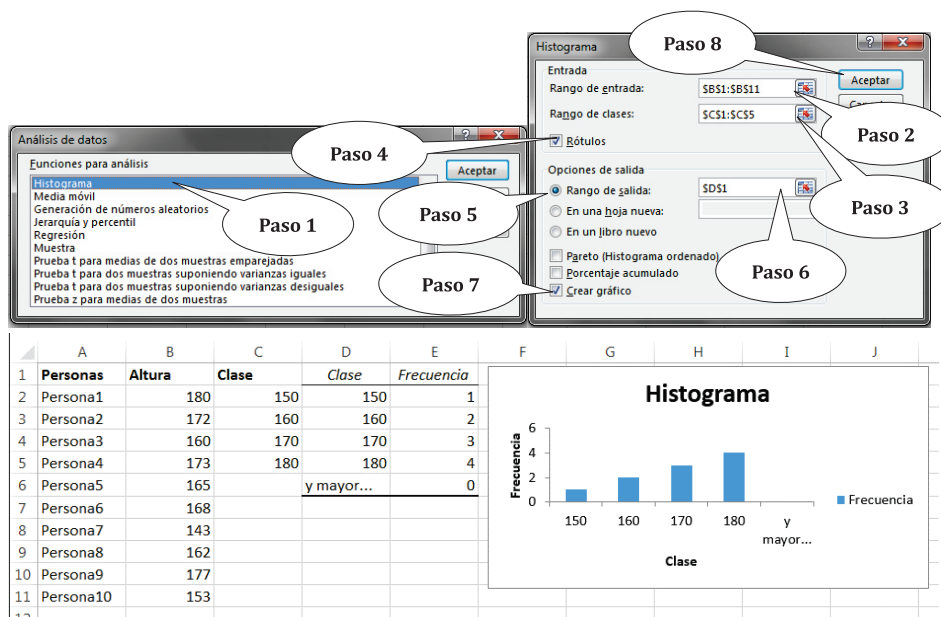


Figura 15. Arriba, ilustración de pasos para generar el histograma; abajo, tabla de resultado del histograma (izquierda) e histograma (derecha).

En la tabla se presentan las frecuencias, que son el resultado de contar la cantidad de valores pertenecientes a cada rango de “Clase”. El histograma nos muestra una distribu-

ción de los datos que aparentemente no se ajusta a la distribución normal, pues las mayores frecuencias están en el extremo derecho y las menores en el extremo izquierdo, representando más bien una distribución asimétrica negativa. Si los datos tuvieran una distribución normal, el histograma sería parecido al ilustrado en la figura 16 A; notemos en la figura 16 B – F los histogramas que nos representan otras distribuciones.

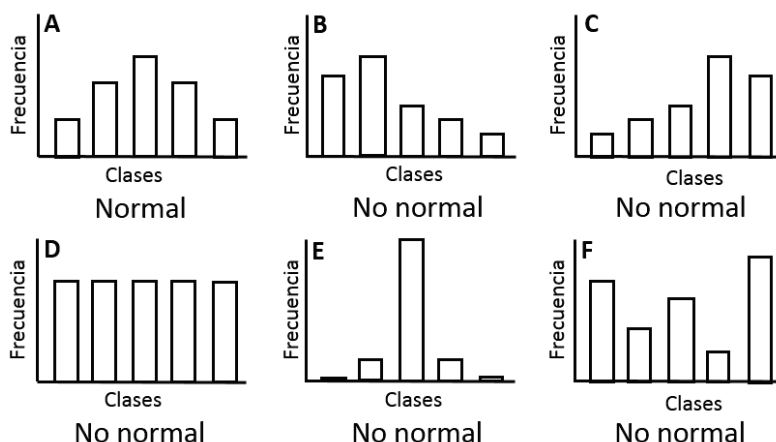


Figura 16. Ilustración para comparar la forma en que un histograma debería visualizarse cuando la distribución de los datos se ajusta y cuando no se ajusta a una distribución normal.

Gráfico Q

El gráfico Q o gráfico cuantil, es una forma común para explorar si un conjunto de datos se distribuyen de forma normal. MS Excel no cuenta con una opción directa para generar este gráfico, de tal forma que el abordaje para su elaboración incluirá procedimientos manuales y procedimientos automáticos combinados.

Utilizaremos el mismo conjunto de datos de altura de personas para ejemplificar la creación del gráfico Q. Primero ordenaremos los datos de forma ascendente, para ellos seleccionamos los datos (sin el encabezado) (Paso 1) (Figura 17) y utilizamos la herramienta “Ordenar” disponible en Datos>Ordenar (Paso 2). El programa nos dará una nota de advertencia con dos opciones: “Ampliar la selección?” o “Continuar con la selección actual”, en cuyo caso seleccionaremos la segunda opción (Paso 3) y hacemos clic en ordenar (Paso 4).

Aparecerá un cuadro de diálogos titulado “Ordenar”, donde definiremos la columna que se va a ordenar, en nuestro caso la “Columna B” (Paso 5). Luego definimos el orden “menor a mayor” (ascendente) (Paso 6). Quitamos el check donde dice “Mis datos tienen en-

Aplicaciones de Estadística Básica

cabezados” ya que no seleccionamos el encabezado (Paso 7). Finalmente presionamos Enter, y así los datos se ordenarán desde 143 (valor mínimo) hasta 180 (valor máximo).

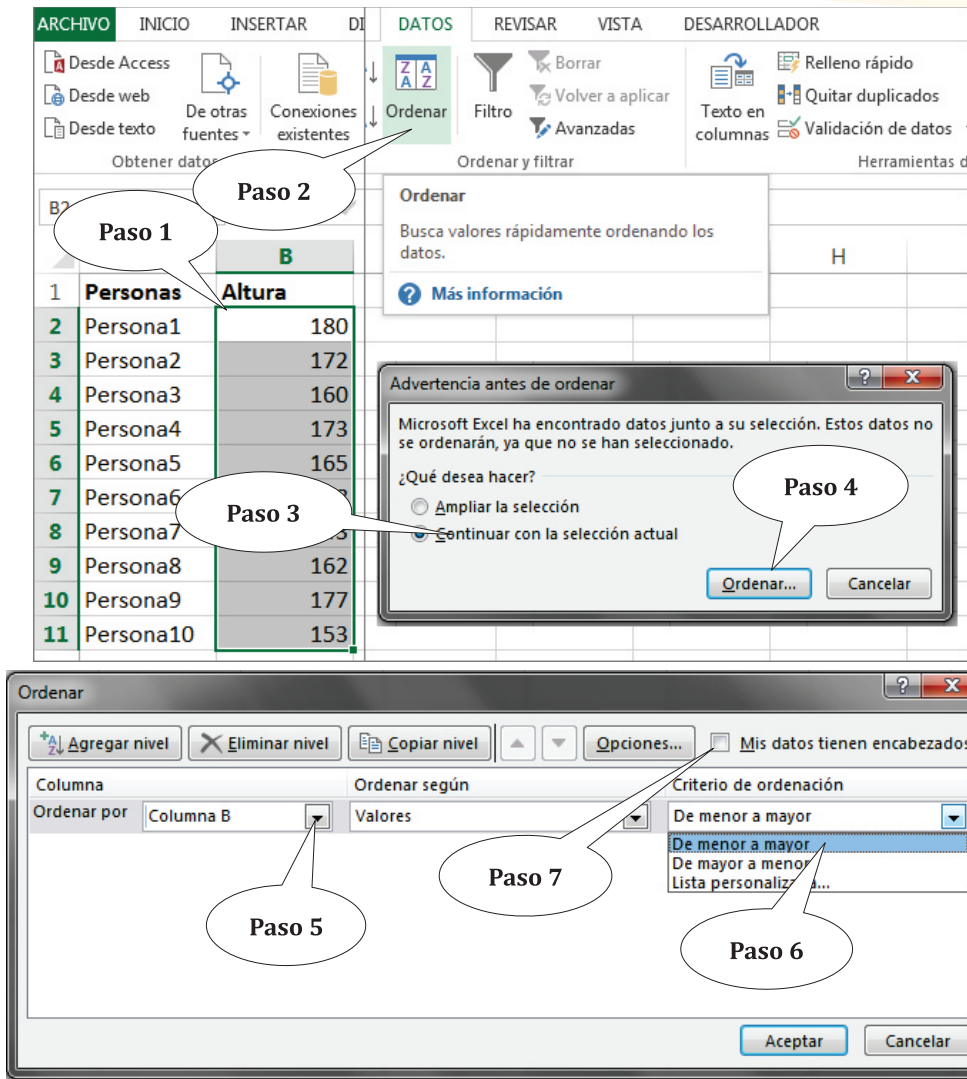


Figura 17. Ilustración de pasos para ordenar el conjunto de datos a utilizar en la creación del gráfico Q.

Una vez ordenados los datos, procederemos a calcular varios valores y variables necesarias para elaborar el gráfico Q. Primero calculamos tres valores: el número de observaciones (#Muestras), el promedio del conjunto de datos y la desviación estándar (DE).

Dado el rango de datos B2:B11, la fórmula para calcular el número de observaciones es =CONTAR(B2:B11), para el promedio es =PROMEDIO(B2:B11) y para la desviación estándar es =DESVEST.M(B2:B11) (Paso 8) (Figura 18).

Seguidamente asignamos una variable formada por números continuos de 1 a 10 (ya que nosotros solo tenemos 10 observaciones), a esta la llamaremos “i” (Paso 9). Con esta variable crearemos otra variable que corresponde a la probabilidad que usaremos para comparar nuestros datos. Esto se logra con la siguiente fórmula:

$$p_i = \frac{i - 0.5}{n}$$

Donde,

p_i = Probabilidad de interés.

n= Número de observaciones.

Sabiendo que el primer valor de “i” (1) está ubicado en la celda C2, la ecuación que calculará la probabilidad para ese número es: =(C2-0.5)/10, lo que es igual a 0.05. Utilizaremos la opción de autorelleno (Figura 4) para que el programa aplique la misma fórmula a los restantes valores de i, creando así una nueva variable a la que llamaremos “=(i-0.5)/n” (Paso 10).

A continuación, calcularemos el valor estandarizado a cada uno de los valores de la variable “=(i-0.5)/n”, dando origen a una nueva variable que llamaremos “Valor Z”. Esto lo logramos utilizando la función: DISTR.NORM.ESTAND.INV() (Paso 11). Para el primer valor de “=(i-0.5)/n” (0.05, celda D2), el valor Z sería: =DISTR.NORM.ESTAND.INV(D2), o sea: -1.644853627. Se utiliza la opción de distribución con el inverso (INV), pues se desea calcular los valores de Z a partir de las probabilidades (se recomienda ampliar conocimiento sobre este tema utilizando otras referencias).

Para finalizar estandarizamos los valores de la variable altura utilizando la ecuación:

$$Z = \frac{x - \mu}{\sigma}$$

Donde,

Z= Valor estandarizado.

x= Observación.

μ = Promedio.

σ = Desviación estándar.

Aplicaciones de Estadística Básica

Si el primer valor es 143 (ubicado en la celda B2), el promedio es 165.3 y la desviación estándar de la muestra es 11.3142, la fórmula aplicada en MS Excel quedaría expresada como: $=(B2-165.3)/11.3142$, lo que resultaría en -1.970974528, que corresponde al valor estandarizado de 165.3. Hacemos el mismo proceso con el resto de los valores de nuestra variable “Altura(cm)” utilizando la opción de autorelleno y de esta forma incluimos la última variable a la que llamaremos “Est_Altura” (Paso 12).

	A	B	C	D	E	F
1	Personas	Altura(cm)	i	(i-0.5)/n	Valor Z	Est_Altura(cm)
2	Persona1	143	1	$=(C2-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D2)$	$=(B2-165.3)/11.3142$
3	Persona2	153	2	$=(C3-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D3)$	$=(B3-165.3)/11.3142$
4	Persona3	160	3	$=(C4-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D4)$	$=(B4-165.3)/11.3142$
5	Persona4	162	4	$=(C5-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D5)$	$=(B5-165.3)/11.3142$
6	Persona5	165	5	$=(C6-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D6)$	$=(B6-165.3)/11.3142$
7	Persona6	168	6	$=(C7-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D7)$	$=(B7-165.3)/11.3142$
8	Persona7	172	7	$=(C8-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D8)$	$=(B8-165.3)/11.3142$
9	Persona8	173	8	$=(C9-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D9)$	$=(B9-165.3)/11.3142$
10	Persona9	177	9	$=(C10-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D10)$	$=(B10-165.3)/11.3142$
11	Persona10	180	10	$=(C11-0.5)/10$	$=DISTR.NORM.ESTAND.INV(D11)$	$=(B11-165.3)/11.3142$
12						
13	# Muestras	$=CONTAR(B2:B11)$				
14	Promedio	$=PROMEDIO(B2:B11)$				
15	DE	$=DESVEST.M(B2:B11)$				

	A	B	C	D	E	F
1	Personas	Altura(cm)	i	(i-0.5)/n	Valor Z	Est_Altura(cm)
2	Persona1	143	1	0.05	-1.644853627	-1.970974528
3	Persona2	153	2	0.15	-1.036433389	-1.087129448
4	Persona3	160	3	0.25	-0.67448975	-0.468437892
5	Persona4	162	4	0.35	-0.385320466	-0.291668876
6	Persona5	165	5	0.45	-0.125661347	-0.026515352
7	Persona6	168	6	0.55	0.125661347	0.238638172
8	Persona7	172	7	0.65	0.385320466	0.592176203
9	Persona8	173	8	0.75	0.67448975	0.680560711
10	Persona9	177	9	0.85	1.036433389	1.034098743
11	Persona10	180	10	0.95	1.644853627	1.299252267
12						
13	# Muestras	10				
14	Promedio	165.3				
15	DE	11.3				

Figura 18. Ilustración de cálculos previos e inclusión de nuevas variables, que se utilizarán en la creación del gráfico Q. Arriba se muestra una hoja de cálculo de MS Excel con las fórmulas y funciones. Abajo se observan los resultados.

Después de calcular valores de las nuevas variables “Valor Z” y “Est_Altura”, procederemos a crear el gráfico Q estableciendo a “Valor Z” en el eje X y a “Est_Altura” en el eje Y. Para ello, seleccionaremos los datos a graficar, sin los encabezados (Paso 13) (Figura 19), y haremos uso de la opción de gráfico de “Dispersión”, que se encuentra en Insertar>Gráficos (Paso 14).

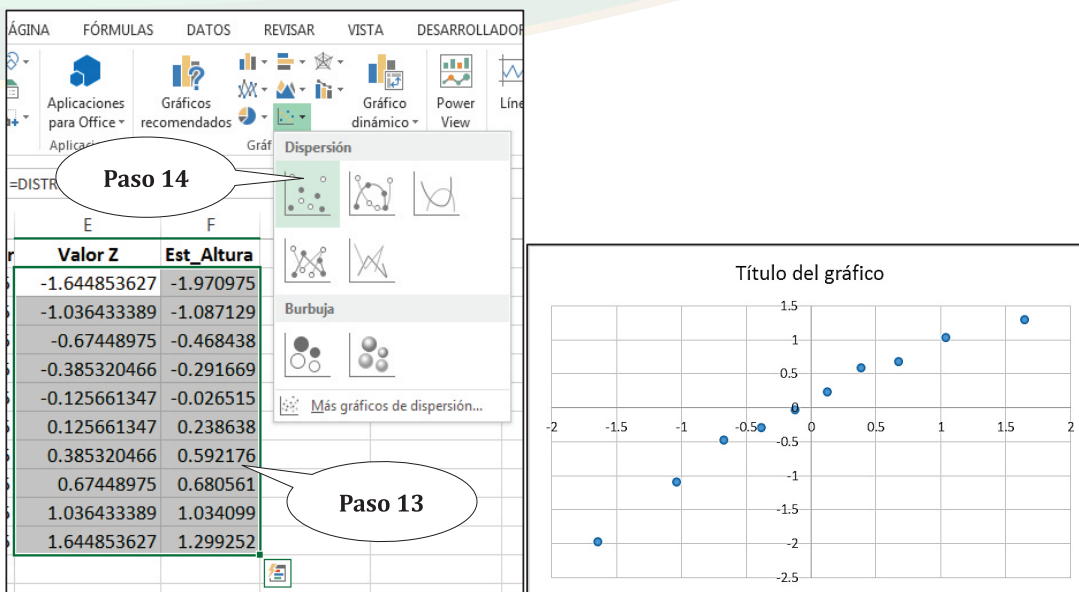


Figura 19. Últimos pasos para creación y presentación del primer boceto de un gráfico Q.

El gráfico Q está parcialmente terminado, y ahora nos tocará hacer algunos arreglos de estética. En primer lugar moveremos los ejes del gráfico, de tal forma que haremos coincidir el eje X con el extremo inferior del gráfico y el eje Y con el extremo izquierdo del mismo. Para lograr este cometido con el eje X, seleccionaremos el eje Y, y haremos clic derecho sobre el mismo, a fin de seleccionar la opción de “Dar formato de eje...” (Paso 15) (Figura 20), aparecerá la paleta de opciones de formato en la parte derecha de la pantalla, donde seleccionaremos la opción “Valor del eje” (Paso 16) y reemplazaremos el valor predeterminado (0.0) por el valor extremo negativo del eje Y o sea -2.5 (Paso 17). En sí, le hemos indicado al programa que el eje X cruce en el valor mínimo del eje Y.

Para dar formato al eje Y se procede de igual forma a como se hizo con el eje X, pero esta vez seleccionaremos la opción “Valor del eje” para el eje X (Paso 18) y en la paleta de opciones reemplazaremos el valor predeterminado (0.0) por el valor extremo negativo del eje X, o sea -2.0 (Paso 19). De esta forma le hemos indicado al programa que el eje Y cruce en el valor mínimo del eje X.

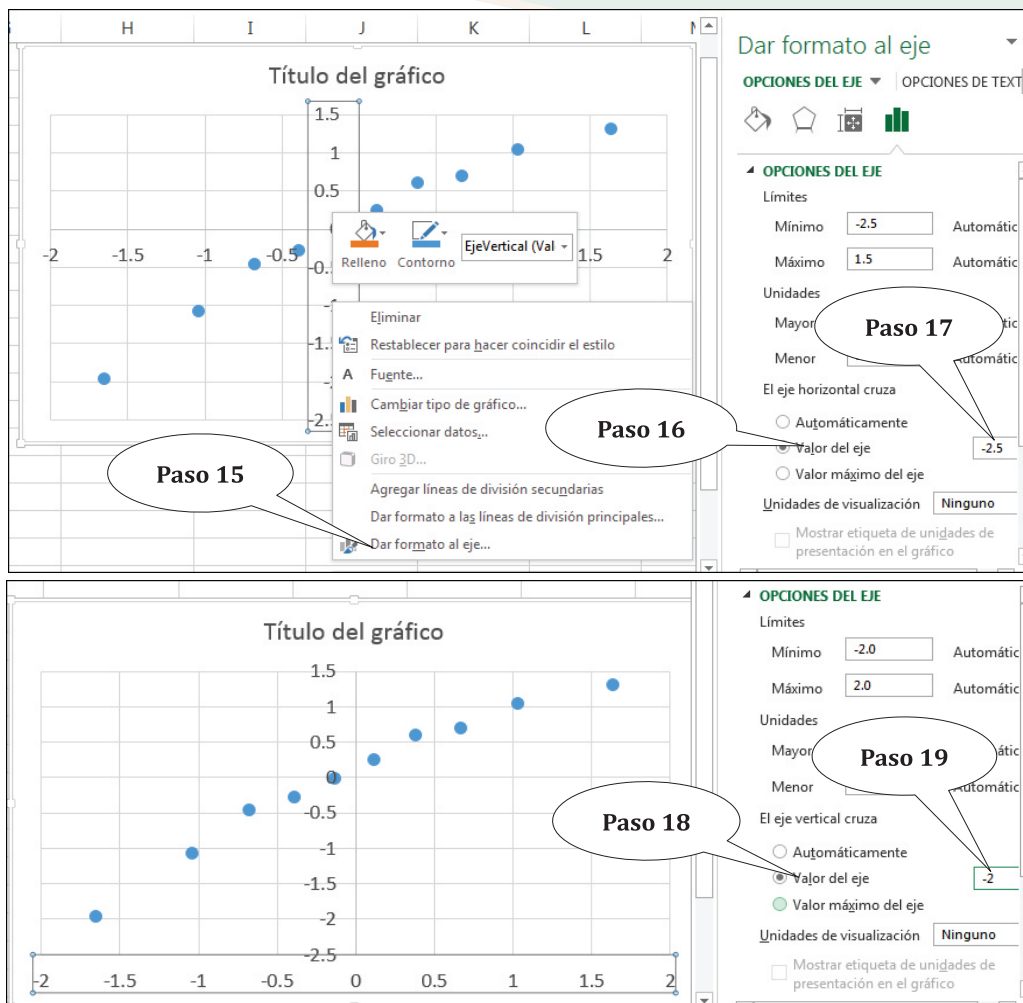


Figura 20. Pasos para personalizar los ejes X y Y del gráfico Q.

Después de posicionar los ejes X y Y en sus lugares, el último paso es insertar una línea de tendencia sobre los puntos del gráfico. Para ello hacemos clic sobre dichos puntos y seleccionamos la opción “Agregar línea de tendencia...” (Paso 20) (Figura 21). Adicionalmente hacemos algunas personalizaciones para mejorar la estética del gráfico, por ejemplo eliminamos el título del gráfico y las rejillas del fondo.

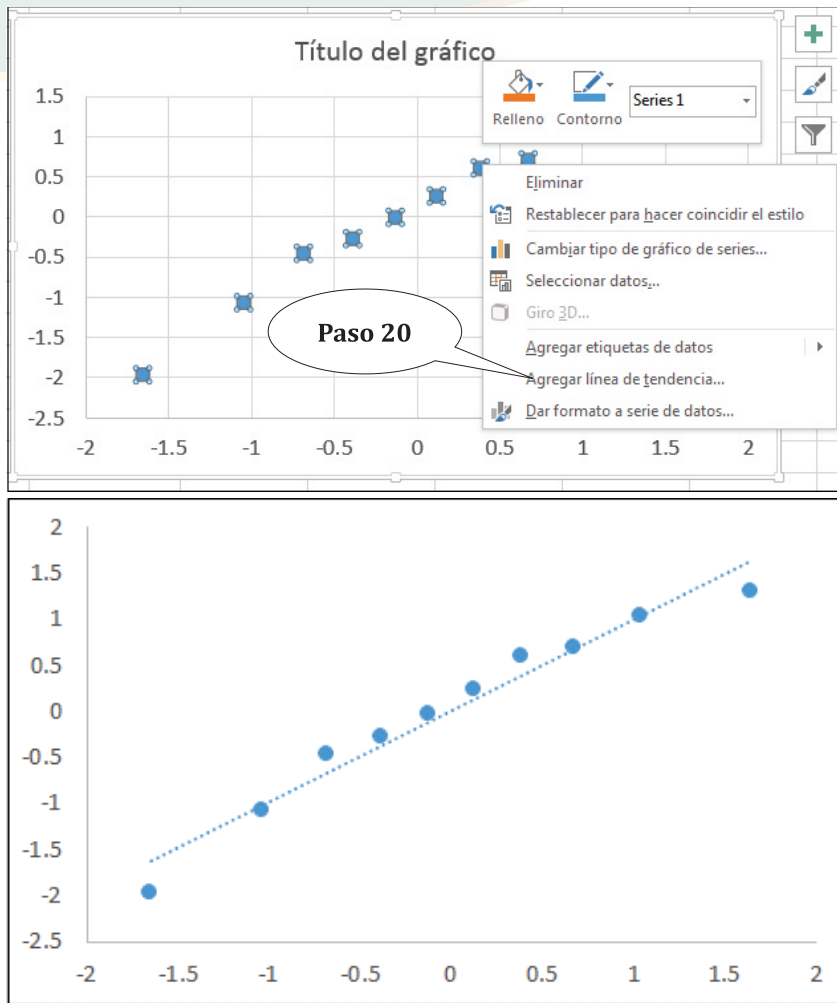


Figura 21. Arriba, se ilustra el paso para agregar la línea de tendencia. Abajo, el gráfico Q finalizado.

En el gráfico Q podemos explorar visualmente la normalidad de los datos, teóricamente los puntos deberían de “más o menos” seguir la línea de tendencia. En la figura 22 se muestra cómo se vería un gráfico Q con datos que se distribuyen de forma normal y las formas que tomaría cuando los datos no siguen una distribución normal. Los datos del ejemplo aparentemente siguen una distribución normal, excepto por dos puntos extremos que se alejan de la línea de tendencia.

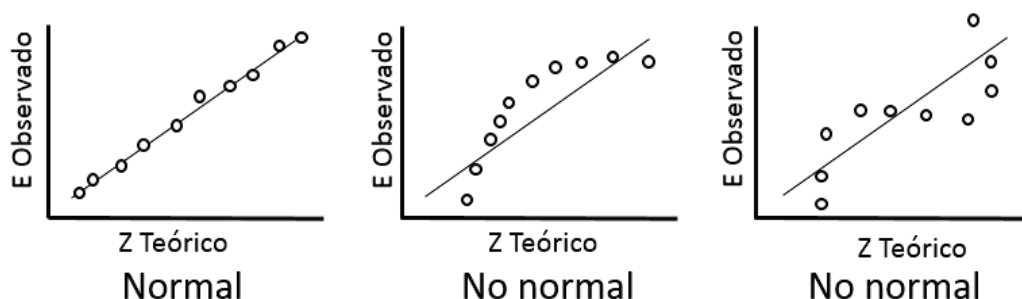


Figura 22. Ilustraciones de tres gráficos Q para visualizar la forma que tomarían los valores (puntos) en correspondencia con la línea de tendencia, cuando se distribuyen de forma normal y cuando no. En el eje X se muestran los valores de Z Teórico y en el eje Y los valores estandarizados de las observaciones.

Métodos numéricos (curtosis y coeficiente de asimetría)

Examinar los datos con las funciones curtosis y coeficiente de asimetría es de vital importancia, pues nos da una idea sólida con respecto a su distribución. Las funciones se encuentran descritas en el cuadro 3, y sus interpretaciones se encuentran en la figura 12 y cuadro 2. Para la curtosis la función es `CURTOSIS()` y para el coeficiente de asimetría la función es `COEFICIENTE.ASIMETRIA()`. En la figura 23 se muestra la aplicación de las dos funciones para los datos de ejemplo.

	A	B
1	Personas	Altura
2	Persona1	180
3	Persona2	172
4	Persona3	160
5	Persona4	173
6	Persona5	165
7	Persona6	168
8	Persona7	143
9	Persona8	162
10	Persona9	177
11	Persona10	153
12		=CURTOSIS(B2:B11)
13		=COEFICIENTE.ASIMETRIA(B2:B11)

Figura 23. Aplicación de las funciones curtosis y coeficiente de asimetría a los datos de altura de personas en cm.

Los resultados de la aplicación de las funciones para los datos mostrados en la figura 23 son:

=CURTOSIS(B2:B11), el valor resultante es: 0.217239979

=COEFICIENTE.ASIMETRIA(B2:B11), el valor resultante es: -0.726899509

Si comparamos los resultados con los valores de referencia del cuadro 2, determinamos que los datos forman una distribución más o menos “Platicúrtica” y evidentemente “Asimétrica negativa”. Con esto tenemos más evidencia para suponer que los datos “no siguen una distribución normal”. Sin embargo, para llegar a confirmaciones precisas, tendremos que aplicar procedimientos inferenciales, como las pruebas de Shapiro-Wilks o Kolmogorov-Smirnov, que no están disponibles como una opción dentro de la herramienta de análisis o a través de una función en MS Excel y cuyo abordaje manual en las hojas de cálculo de MS Excel es un tanto tedioso. Para aplicar estas pruebas se recomienda instalar algún complemento de estadística en MS Excel que incluya dichas pruebas o utilizar el programa R.

Transformaciones

Cuando determinamos que los datos no siguen una distribución normal, no es recomendado aplicar alguna prueba paramétrica cuya premisa (no la única) es que los datos deben seguir dicha distribución. Una forma de ajustar los conjuntos de datos que no se ajustan a una distribución normal, es transformándolos. Hay muchas formas de transformar los datos, en la figura 24 se muestran algunas de ellas: Normal, referido a la normalización o estandarización de los datos, se realiza sustrayendo la media a cada dato y dividiendo el resultado entre la desviación estándar muestral; Ln, o logaritmo neperiano; Log10, o logaritmo base 10; Log2, o logaritmo base 2; RaizCuad., que representa a la raíz cuadrada de los valores; Potencia, en el que se eleva a una potencia deseada; Inverso, en el que se divide uno entre cada valor.

	A	B	C	D	E	F	G	H	I
1	Nº Punto	pH	Normal	Ln	Log10	Log2	RaizCuad	Potencia	Inverso
2	Punto1	4.7	=(B2-5.08)/0.53	=LN(B2)	=LOG10(B2)	=LOG(B2,2)	=RAIZ(B2)	=POTENCIA(B2,5)	=1/B2
3	Punto2	5.3	=(B3-5.08)/0.53	=LN(B3)	=LOG10(B3)	=LOG(B3,2)	=RAIZ(B3)	=POTENCIA(B3,5)	=1/B3
4	Punto3	5.9	=(B4-5.08)/0.53	=LN(B4)	=LOG10(B4)	=LOG(B4,2)	=RAIZ(B4)	=POTENCIA(B4,5)	=1/B4
5	Punto4	4.9	=(B5-5.08)/0.53	=LN(B5)	=LOG10(B5)	=LOG(B5,2)	=RAIZ(B5)	=POTENCIA(B5,5)	=1/B5
6	Punto5	4.6	=(B6-5.08)/0.53	=LN(B6)	=LOG10(B6)	=LOG(B6,2)	=RAIZ(B6)	=POTENCIA(B6,5)	=1/B6
7	Media	=PROMEDIO(B2:B6)							
8	DE	=DESVEST.M(B2:B6)							

	A	B	C	D	E	F	G	H	I
1	Nº Punto	pH	Normal	Ln	Log10	Log2	RaizCuad	Potencia	Inverso
2	Punto1	4.7	-0.71698113	1.54756251	0.67209786	2.23266076	2.16794834	2293.45007000	0.21276596
3	Punto2	5.3	0.41509434	1.66770682	0.72427587	2.40599236	2.30217289	4181.95493000	0.18867925
4	Punto3	5.9	1.54716981	1.77495235	0.77085201	2.56071495	2.42899156	7149.24299000	0.16949153
5	Punto4	4.9	-0.33962264	1.58923521	0.69019608	2.29278175	2.21359436	2824.75249000	0.20408163
6	Punto5	4.6	-0.90566038	1.52605630	0.66275783	2.20163386	2.14476106	2059.62976000	0.21739130
7	Media	5.08							
8	DE	0.53103672							

Aplicaciones de Estadística Básica

Figura 24. Ilustración de una hoja de cálculo de MS Excel con los datos de pH de suelo, demostrando siete formas de transformar los datos. Normal= Estandarización de los valores; Ln= Logaritmo natural; Log10= Logaritmo base 10; Log2= Logaritmo base 2; RaizCuad= Raíz cuadrada; Potencia= Uso de potencias (para este ejemplo se elevó a la potencia 5, también se puede usar “=B2^5”) e Inverso= conversión a 1/Valores. Notar el cálculo de la media y la desviación estándar como requisito a usarse en algunas transformaciones.

Estadística inferencial

En la estadística inferencial se pretende aplicar una prueba para generar un valor probabilístico para “rechazar” o “fallar en rechazar” una hipótesis. Cada prueba tiene su propia hipótesis y en este acápite se presentarán dichas hipótesis. Con la salvedad que no será interés de este libro explicar la naturaleza y génesis de las hipótesis, para lo cual recomendamos mayor documentación en algún libro de estadística. Nuestro principal interés es el ilustrar cómo se utiliza MS Excel para aplicar dichas pruebas. Este capítulo está dividido en cuatro temas: ¿Qué es “p”? ¿Probabilidad de qué? ¿Alfa? ¿Hipótesis?; comparación de proporciones y frecuencias; comparación de medias; y relaciones entre variables.

¿Qué es “p”? ¿Probabilidad de qué? ¿Alfa? ¿Hipótesis?

En esta sección se pretenden responder, de manera general, algunas preguntas comunes que agobian a los novicios en estadística, las cuales son: ¿Qué es “p”? ¿Probabilidad de qué? ¿Alfa? ¿Hipótesis? No se pretende brindar una cátedra completa de estos temas, pues existe una vasta literatura que ofrece respuestas para estas preguntas, pero se hará énfasis en algunas explicaciones específicas para comprender la estadística inferencial. El valor de “p” es sumamente importante en la estadística inferencial, pues nos ayuda a rechazar o fallar de rechazar (o sea aceptar) una hipótesis. De acá en adelante determinar el valor de “p” será crucial en todas las pruebas estadísticas que se ejemplificarán, utilizando tanto Microsoft Excel como R. La “p” es la abreviación de “probabilidad”, y denota en específico a la probabilidad de encontrar valores extremos en una muestra. Para entender a qué se refiere esta probabilidad tendremos que introducir un par de conceptos más, uno es el concepto de “distribución normal” y el otro es el concepto de “nivel de significancia” o “ α ”.

Es necesario aclarar que la explicación del valor de “p” en este libro excluye las fundamentaciones formales y tradicionales de un libro de texto de estadística. En este contexto, se trata de proporcionar explicaciones sencillas y accesibles, para un lector que no necesariamente es, ni será, profesional en estadísticas; sino, para un lector que necesita utilizar estadísticas a fin de resolver problemas y tomar decisiones propias de su contexto profesional.

En el acápite llamado “Glosario de términos relacionados con la clave” se introdujo rápidamente el concepto de “distribución normal”, luego retomamos de nuevo el tema en la sección llamada “La distribución normal”, ahora volveremos a mencionar este término para detallar más algunos aspectos particulares relacionados con las pruebas inferenciales. Cuando tenemos un conjunto de datos proveniente de medidas (por ejemplo), estos pueden distribuirse de forma normal, o sea, adquiriendo la forma de una campana (campana de Gauss). En la figura 25 A se ofrece una ejemplificación, en la cual se midió la estatura de un número determinado de personas, luego esas estaturas se agruparon en rangos (por ejemplo, de 5 en 5 o de 10 en 10) y se contó el número de observaciones (personas) que pertenecían a cada rango de estatura. Como resultado se obtiene un gráfico con forma de campana, donde en la parte central se encuentra la mayoría de los individuos observados, quienes tienen estatura promedio; al lado izquierdo están los individuos observados de estatura baja (en relación al promedio) y por ende son una minoría; al lado derecho se agrupan los individuos observados de estatura alta (en relación al promedio) los cuales también son una minoría.

Esa distribución en la cual “la mayoría está en el centro y pocos a los lados” hace que el gráfico adquiera forma de campana. Teóricamente se ha designado que las observaciones que no son casos extremos conforman el 95% del área de esa campana y el restante 5% conforman los casos extremos; o sea (siguiendo con el ejemplo) ese 5% corresponde a las personas que son muy bajas o son muy altas y por ende el 2.5% del área de la campana estará representada por la minoría extremo de estatura baja localizada a la izquierda y el otro 2.5% del área de la campana estará representada por la minoría extremo de estatura alta, localizada a la derecha (Figura 25 B). Entonces en la campana cada uno de esos 2.5% representan las “áreas críticas” de la campana, o sea el área que denota cuando una observación es considerada extremo, en el caso del ejemplo, o muy pequeño o muy grande (Figura 25 C).

A ese 5% de observaciones que son extremos se les denomina alfa (α) o “nivel de significancia”, el cual también se suele representar como 0.05 o sea 5/100. El alfa se contrasta con el valor de “p”, cuando “p” es menor que “ α ” se rechaza la hipótesis nula y, por lo tanto, se acepta la alternativa, y cuando “p” es mayor o igual a “ α ” se falla en rechazar la hipótesis nula. En este contexto entonces cabe introducir también el concepto de “hipótesis”.

En estadística inferencial, una hipótesis es un hecho que se asume verdadero mientras no haya evidencias para probar lo contrario. Desde este punto de vista, y de manera general, se derivan dos hipótesis, una hipótesis llamada “nula” simbolizada por “ H_0 ” y una hipótesis llamada “alternativa” simbolizada por “ H_1 ” o también como “ H_a ”. Ambas hipótesis son obvias y lógicamente contrastantes. La hipótesis nula denota una “igualdad”, un “no cambio”, un “no efecto”; la hipótesis alternativa denota una “desigualdad”,

Aplicaciones de Estadística Básica

un “cambio”, un “efecto”. Por ejemplo, la expresión $H_0: \mu_1 = \mu_2$ representa la hipótesis nula en que la media 1 (μ_1) es igual a la media 2 (μ_2); opuestamente la expresión $H_1: \mu_1 \neq \mu_2$ representa la hipótesis alternativa en que la media 1 (μ_1) es diferente a la media 2 (μ_2). Para seleccionar la hipótesis se utiliza, entonces la comparación del valor de “p” y el de “ α ”.

Con lo anterior en mente, podemos deducir que el 5% que representa “ α ” corresponde a la zona de “rechazo de H_0 ” y el restante 95% corresponde a la zona de “fallar en rechazar H_0 ” (Figura 25 D). Si contextualizamos al ejemplo de las estaturas, una persona cuyo tamaño se corresponda a cualquier tamaño dentro de la zona de 95%, se considerará una persona de estatura no extrema, pero si hay una persona cuyo tamaño es muy pequeño y corresponde al tamaño dentro de la zona de 2.5% a la izquierda, se considerará una persona extremadamente pequeña; de igual forma, una persona cuyo tamaño es muy grande y corresponde al tamaño dentro de la zona de 2.5% a la derecha, se considerará una persona extremadamente grande. De igual forma si la diferencia entre dos conjuntos de datos cae en la zona de 95%, se considera una diferencia no significativa, mientras que si esa diferencia cae en la zona de 5% se considera significativa.

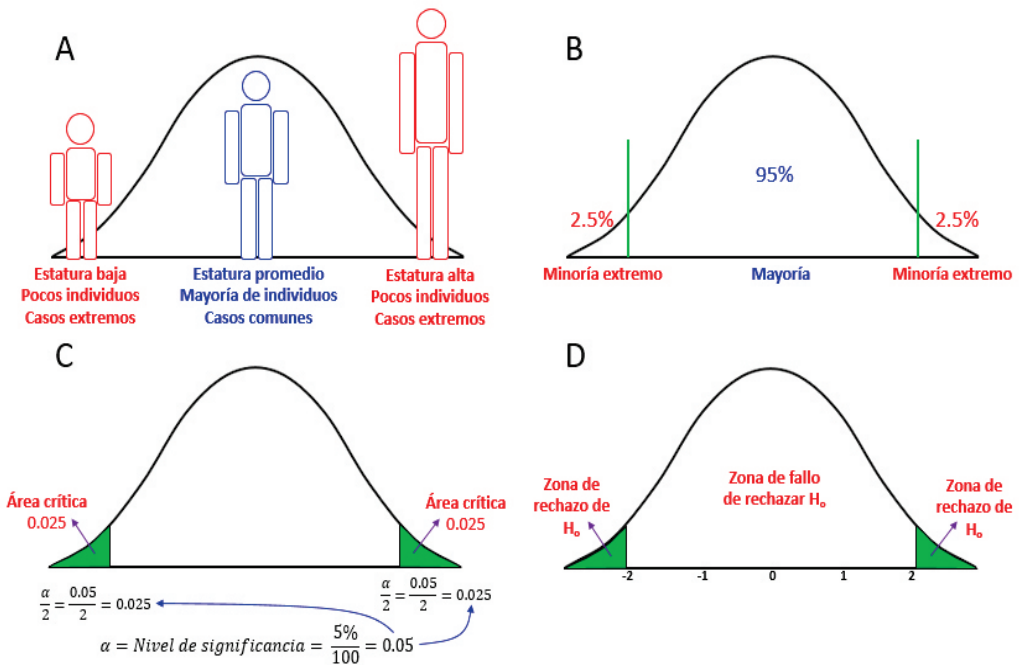


Figura 25. Representación gráfica para ilustrar los conceptos de normalidad, nivel de significancia e hipótesis.

Para entender un poco más la dinámica entre “p” y “ α ” haremos una ampliación de una de las zonas extremos de la campana de Gauss (Figura 26). La frontera entre la zona de rechazo de H_0 (área verde) y la zona de fallar a rechazar H_0 (área blanca) la marca un valor que se llama “valor crítico”. Si el área de “p” (área beige claro) sobrepasa ese valor crítico, o sea es más grande que el área en verde ($\alpha = 0.025$), se falla en rechazar H_0 y se concluiría “igualdad” o “no significancia”, pues el valor del estadístico que define el límite izquierdo de “p” cayó en la zona de “fallar a rechazar” (Figura 26 A).

Por otra parte, si el área de “p” (área beige claro) no sobrepasa ese valor crítico, o sea es más pequeño que el área en verde ($\alpha = 0.025$), entonces se rechaza H_0 y se asume H_1 la cual considera “desigualdad” o “significancia” pues el valor del estadístico que define el límite izquierdo de “p” cayó en la zona de “rechazo” (Figura 26 B). En el ejemplo de la figura 26 A se observa que “p” es mayor que “ α ”, dado que $p = 0.13 > \alpha = 0.025$; por tanto, se falla en rechazar H_0 . En la figura 26 B, “p” es menor que “ α ”, dado que $p = 0.012 < \alpha = 0.025$; por tanto, se rechaza H_0 y se acepta H_1 .

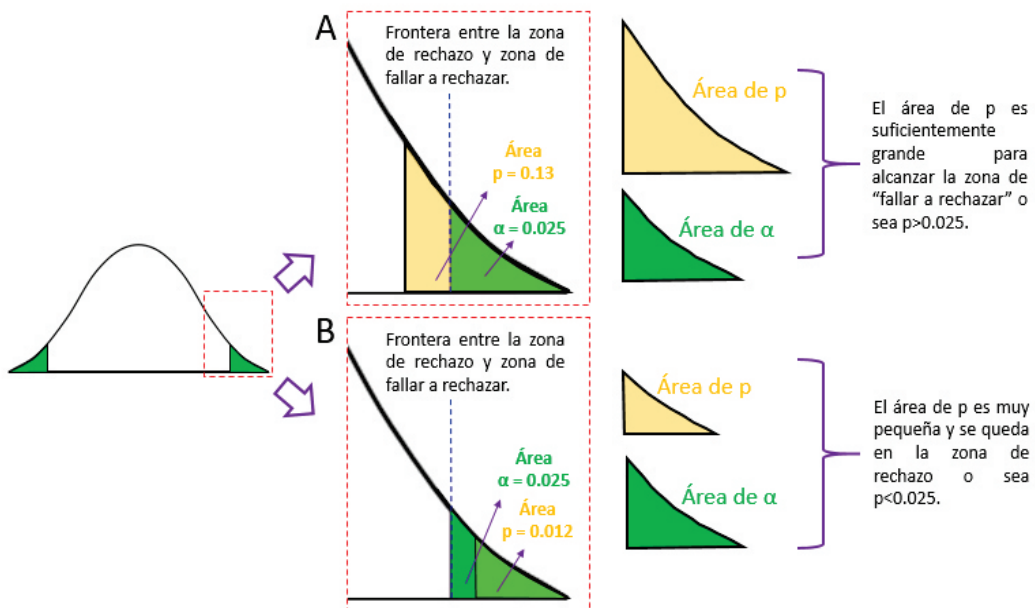


Figura 26. Representación gráfica de la interacción entre el valor de “p” y el valor de “ α ”, haciendo el valor de este último 0.025 para este ejemplo (se representa con 0.05 en la figura 27). Notar el cambio de color en la superposición de los colores beige claro y verde, el área superpuesta se muestra en un color verde claro.

Aplicaciones de Estadística Básica

Cuando la prueba estadística designa una comparación y asume una hipótesis nula con una connotación de igualdad, por ejemplo $H_0: \mu_1 = \mu_2$ (μ =media) el área de rechazo denotada por $\alpha = 0.05$ se divide entre las dos colas de la campana, o sea conforma las áreas en verde de la campana en la figura 27 A. Si a priori se conoce que el valor de un parámetro en una observación es mayor que el otro, la hipótesis nula siempre sigue denotando igualdad ($H_0: \mu_1 = \mu_2$), pero la hipótesis alternativa representa la dirección de la desigualdad, por ejemplo $H_1: \mu_1 > \mu_2$. Para este caso, el área de rechazo representado por $\alpha = 0.05$ se concentra en la cola derecha de la campana (Figura 27 B). Si a priori se conoce que el valor de una medida en una observación es menor que el otro, por ejemplo $H_1: \mu_1 < \mu_2$, el área de rechazo denotado por $\alpha = 0.05$ se concentra en la cola izquierda de la campana (Figura 27 C).

Tanto para la prueba de cola derecha como de cola izquierda, el contraste entre “p” y “ α ” siempre va a tener la misma dinámica, si $p > 0.05$ se falla en rechazar H_0 y se concluye “no significancia” y si $p < 0.05$ se rechazar H_0 , se acepta H_1 y se concluye “significancia”. Una diferencia entre los dos tipos de pruebas es que las pruebas de colas derechas tendrán los valores críticos y valores de los estadísticos con signos positivos y las pruebas de colas izquierdas, tendrán los valores críticos y valores de los estadísticos con signos negativos (Figura 27 D – G).

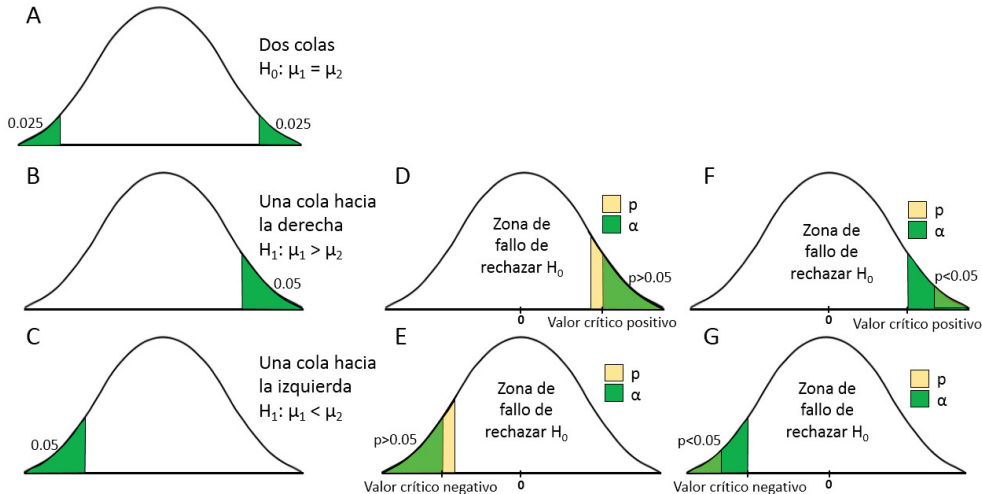


Figura 27. Representación gráfica de la interacción entre el valor de “p” y el valor de “ α ”, en pruebas de dos colas, una cola hacia la derecha y una cola hacia la izquierda. Notar el cambio de color en la superposición de los colores beige claro y verde, las áreas superpuestas se muestran en un color verde claro.

Es bueno señalar que el valor de “ α ” puede ser definido por el usuario, en dependencia de la precisión con la que quiere probar las hipótesis. Los niveles que comúnmente se usan son: 1% (0.01), 5% (0.05) o 10% (0.1). Por ejemplo, en experimentos farmacéuticos con variables controladas el nivel de significancia a usarse puede ser 0.01, o sea el chance de error es muy pequeño; en muestreos biológicos sin variables controladas el nivel de precisión es más holgado y puede usarse 0.05 o 0.1 (menos usual). En este escrito para todos los ejercicios utilizaremos el nivel de significancia igual a 0.05.

La prueba de hipótesis, en resumidas cuentas, sirve para tomar decisiones, y a como vimos, dos decisiones se pueden tomar: “rechazar H_0 ” o “fallar en rechazar H_0 ”; sin embargo, tenemos que tomar en cuenta que en toda toma de decisión también se puede incurrir en errores. En este sentido, hay dos errores que se pueden cometer: “rechazar H_0 cuando se tiene que aceptar” o “aceptar H_0 cuando se tiene que rechazar”, a esto se les llama los errores de “Tipo I” y “Tipo II” respectivamente.

Aunque no es el objetivo de este libro detallar sobre los dos tipos de errores, hay mucha literatura que podría abordar el tema a detalle, haremos una pequeña descripción de cada uno. El error de tipo I se comete al rechazar H_0 cuando es verdadera (o sea, se rechaza cuando no se debía rechazar). Este error se representa por α , o sea, por el mismo nivel de significancia. El error Tipo I es considerado “muy serio” pues cuando ocurre básicamente se está reportando un efecto que no existe, por ejemplo, reportar que un tratamiento tuvo efecto, cuando realmente no lo tuvo. Como el investigador es quien selecciona el nivel de significancia, él mismo sería el culpable de ese error. En general este error ocurre por aplicar diseños experimentales o de muestreo con ciertos sesgos en la colecta de los datos. Reducir α puede evitar el error de “Tipo I”, pero puede incrementar el error de “Tipo II”.

El error de Tipo II se comete cuando fallamos de rechazar, o sea la aceptamos H_0 cuando realmente es falsa (o sea, no se rechaza cuando se debería rechazar). Este error se representa por β y es considerado “menos serio”, pues cuando ocurre lo que se está reportando, es que no existe un efecto, cuando realmente lo existe, en cuyo caso el investigador pasa por inadvertido algún cambio que estaba esperando, por ejemplo, no reportar que un tratamiento tuvo efecto cuando realmente lo tuvo. El error Tipo II se produce al tener un tamaño de muestra muy pequeña o cuando la variabilidad de la población examinada es muy grande.

Para reducir la probabilidad de cometer, tanto el error de Tipo I como el error de Tipo II, la clásica recomendación (aunque no la única) es aumentar el tamaño de la muestra, lo cual además incrementar el poder de las pruebas estadísticas.

Aplicaciones de Estadística Básica

Comparación de proporciones y frecuencias

Las comparaciones de proporciones y frecuencias son muy comunes, en especial cuando se trabaja con información cualitativa. Una proporción es un valor relativo de un dato en relación al total del conjunto de datos al que pertenece. En datos de frecuencia, una proporción será igual al número de frecuencias de una característica dividido entre el total de frecuencias de todas las características. Por ejemplo: Se le pregunta a 50 persona si están de acuerdo con un plan de conservación y 42 responden que SÍ y 8 responden que NO, entonces las proporciones de respuestas de “SÍ” o “NO” serían:

$$\text{SÍ} = \frac{42}{50} = 0.84$$

$$\text{NO} = \frac{8}{50} = 0.16$$

Toda proporción puede ser expresada en porcentaje si se multiplica por 100, de tal forma que en el ejemplo, en 84% de las personas dijeron “SÍ” y el 16% dijeron “NO”. La sumatoria de las proporciones de un conjunto de datos debe ser 1; así $0.84 + 0.16 = 1$.

La frecuencia es el número de veces que se cuenta un objeto, evento, organismo o número. Ejemplo: Numero de huevos en un nido; número de plantas dañadas por un hongo; número de personas de tamaño entre 1.5 y 1.8 metros, número de veces que una persona responde “SÍ” o “NO” en una encuesta, etc.

A continuación se explicarán los procedimientos para aplicar prueba de una proporción, donde se estudiará cómo se compara una proporción proveniente de un dato experimental o de muestreo con una teórica; prueba de dos proporciones donde se abordará la comparación de dos proporciones proveniente de datos experimentales o muestrales; bondad de ajuste, en la cual se medirá el ajuste de frecuencias con proporciones preestablecidas; pruebas de independencia aplicadas específicamente a frecuencia, donde se prueba la dependencia o independencia entre 2×2 condiciones, o entre $R \times C$ condiciones (R = más de dos condiciones en las filas, C = más de dos condiciones en las columnas).

Para la mayoría de estas pruebas, MS Excel no cuenta con una opción o función automática, sino que el abordaje se explicará de forma manual con las hojas de cálculo y algunas funciones; sin embargo, este abordaje no es tan tedioso (relativamente). Para un mejor entendimiento por parte del lector, los procedimientos se realizarán paso a paso.

Prueba de una proporción

Con esta prueba se pretende comparar una proporción, que se ha originado mediante muestreo o de forma experimental (observada) con otra proporción preestablecida. La prueba de una proporción asume las siguientes hipótesis:

$$H_0: \hat{p} = p$$

$H_1: \hat{p} \neq p$; también llamada de dos colas.

$H_1: \hat{p} > p$; también llamada de una cola hacia la derecha.

$H_1: \hat{p} < p$; también llamada de una cola hacia la izquierda.

Donde,

\hat{p} = Proporción observada.

p = Proporción teórica.

En dependencia del tipo de prueba de proporciones (una o dos colas), así se utilizan las fórmulas y funciones en MS Excel, el cuadro 5 presenta cada situación.

Cuadro 5. Tipo colas y las fórmulas y funciones asociadas para lograr los cálculos.

COMPARACIONES	FÓRMULA
Una cola - hacia la derecha	=1-DISTR.NORM.ESTAND.N(z,VERDADERO)
Una cola - hacia la izquierda	=DISTR.NORM.ESTAND.N(z,VERDADERO)
Dos colas	=2*MIN (Una cola - hacia la izquierda;Una cola - hacia la derecha)

Si el argumento acumulado es VERDADERO, DISTR.NORM.ESTAND.N devuelve la función de distribución acumulativa; si es FALSO, devuelve la función de densidad de probabilidad.

z=Valor del estadístico Z.

Para ejemplificar la prueba de una proporción para una cola hacia la derecha, se utilizará la siguiente situación hipotética: En un estudio de sanidad de peces en un río, se presume que el número de peces afectados por un parásito es mayor al 10%. Para probar si este valor es correcto, se realiza un muestreo aleatorio de peces donde se revisan 756 peces, de los cuales 134 estaban afectados por el parásito. Se pretende determinar si la proporción de los datos muestreados, realmente coincide con la hipótesis $>10\%$ asumida.

Aplicaciones de Estadística Básica

Aproximaremos esta distribución binomial (peces afectados versus peces no afectados) utilizando la distribución normal. Para ello, nuestros datos tienen que llenar dos requisitos:

$$n \times p \geq 10 \text{ y } n \times (1 - p) \geq 10$$

Donde,

n = Número de observaciones.

p = Proporción teórica, para el ejemplo este valor es igual al 10% o 0.1 (forma de proporción).

Para nuestro ejemplo el resultado es:

$$756 \times 0.1 = 75.6 \text{ y } 756 \times (1 - 0.1) = 680.4$$

Ambos valores resultantes (75.6 y 680.4) son mayor que 10, por lo que se cumplen los criterios para utilizar la distribución normal como aproximación.

El uso de una distribución normal para aproximar una distribución binomial, necesita una corrección llamada “Corrección de Continuidad”, que consiste en restar o sumar 0.5 al valor de la característica de interés (número de peces afectados). Si la dirección de la prueba es hacia la derecha restaríamos 0.5; si fuese hacia la izquierda sumaríamos 0.5 y si fuera de dos colas, haríamos la resta para la cola derecha, la suma para la cola izquierda y seleccionaríamos el menor valor de “ p ” resultante de ambas correcciones, para finalmente multiplicarlo por 2.

En principio, calculamos la proporción de la muestra dividiendo el número de individuos afectados entre el total de individuos muestreados o sea $\frac{134}{756} = 0.177248677$. Sin embargo, para la prueba utilizaremos la proporción con la corrección (- 0.5), o sea $\frac{134-0.5}{756} = \frac{133.5}{756} = 0.176587302$ (Paso 1), seguidamente calculamos el valor del estadístico Z utilizando la fórmula:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1 - p)}{n}}}$$

Donde,

Z = Estadístico.

\hat{p} = Proporción muestreada, en la figura 28 este valor se encuentra en la celda B6.

p = Proporción teórica, para el ejemplo este valor es igual al 10% o 0.1 (celda B5).

n = Número de observaciones (celda B2).

Entonces expresando la fórmula anterior en formato de MS Excel, el cálculo del estadístico Z lo realizaríamos con la expresión: $=(B6-B5)/RAIZ((B5*(1-B5))/B2)$ (Paso 2), lo que equivale a 7.019342136. A continuación calculamos el valor de la probabilidad (p), sabiendo con anticipación que es una cola hacia la derecha, por lo que se implementa la fórmula respectiva, según el cuadro 5.

Si el valor del estadístico Z se ubica en la celda B7, la fórmula en formato de MS Excel para el cálculo de la probabilidad sería: $=1-DISTR.NORM.ESTAND.N(B7,VERDADERO)$ (Paso 3), cuyo resultado es 1.11455E-12, lo que es equivalente a 1.11×10^{-12} , o sea 0.00000000000111. Dado $\alpha = 0.05$, el valor de p es mucho menor que 0.05, por lo que se rechaza la hipótesis nula y se concluye que realmente el número de peces afectados por un parásito es mayor al 10%.

	A	B		A	B
1	Peces infectados	=134-0.5	1	Peces infectados	133.5
2	Muestra (n)	756	2	Muestra (n)	756
3	Hipótesis 0	p=0.1	3	Hipótesis 0	p=0.1
4	Hipótesis A	p>0.1	4	Hipótesis A	p>0.1
5	p teórico	0.1	5	p teórico	0.1
6	p muestreado	=B1/B2	6	p muestreado	0.176587302
7	Valor Z	=(B6-B5)/RAIZ((B5*(1-B5))/B2)	7	Valor Z	7.019342136
8	Valor P	=1-DISTR.NORM.ESTAND.N(B7,VERDADERO)	8	Valor P	1.11455E-12

Figura 28. Pasos para aplicar la prueba de una proporción para una cola hacia la derecha, utilizando abordaje manual y algunas funciones. A la izquierda se presentan las fórmulas; a la derecha los valores resultantes.

Para ejemplificar la prueba de una proporción para una cola hacia la izquierda, se utilizará la siguiente situación hipotética: En un estudio de mortalidad de plantas de una especie de árbol, sembrada en un plan de reforestación, se presume que la mortalidad de las mismas es menor al 20%. Pasado un lapso de tiempo después de haberlas sembrado, y para confirmar que la mortalidad realmente era menor al 20%, se realizó un muestreo de 1393 plantas y se verificó que 102 estaban muertas. De tal forma que es interés del estudio el comparar la proporción muestreada, con la teórica.

Aplicaciones de Estadística Básica

Antes de iniciar, evaluaremos si los datos cumplen los dos requisitos, para aproximar la distribución binomial utilizando la distribución normal:

$$n \times p \geq 10 \text{ y } n \times (1-p) \geq 10$$

$$1393 \times 0.2 = 278.6 \text{ y } 1393 \times (1-0.2) = 1114.4$$

Ambos valores resultantes (278.6 y 1114.4) son mayor que 10, por lo que se cumplen los criterios para utilizar la distribución normal como aproximación.

Para iniciar la prueba, calculamos la proporción de la muestra dividiendo el número de individuos afectados entre el total de individuos muestreados, o sea $\frac{102}{1393} = 0.073223259$. Sin embargo, para la prueba utilizaremos la proporción con la corrección (+ 0.5, cola izquierda), o sea $\frac{102 + 0.5}{1393} = \frac{102.5}{1393} = 0.073582197$ (celda B6) (Paso 1) (Figura 29), seguidamente calculamos el valor del estadístico con la fórmula de Z en MS Excel: $=(B6-B5)/\text{RAIZ}((B5*(1-B5))/B2)$ (en B2 se encuentra el número de observaciones y en B5 la proporción teórica) (Paso 2), lo cual se obtiene como resultado -11.79570239 (celda B7), luego determinamos el valor de la probabilidad (p) mediante la fórmula: $=\text{DISTR.NORM.ESTAND.N}(B7,\text{VERDADERO})$ (Paso 3), resultando el valor de "p" igual a 2.0538E-32, lo que es equivalente a 2.05×10^{-32} (en R se ha presentado el valor aproximado: $<2.2\text{E-}16$). Dado $\alpha = 0.05$, el valor de p es extremadamente menor que 0.05, por lo que rechazamos la hipótesis nula y se concluye que realmente el número de plantas muertas es menor que el 20%.

	A	B	C
1	Plantas muertas	=102+0.5	
2	Muestra (n)	1393	
3	Hipótesis 0	p=0.2	
4	Hipótesis A	p<0.2	
5	p teórico	0.2	
6	p muestreado	=B1/B2	
7	Valor Z	=(B6-B5)/RAIZ((B5*(1-B5))/B2)	
8	Valor P	=DISTR.NORM.ESTAND.N(B7,VERDADERO)	

	A	B
1	Plantas muertas	102.5
2	Muestra (n)	1393
3	Hipótesis 0	p=0.2
4	Hipótesis A	p<0.2
5	p teórico	0.2
6	p muestreado	0.073582197
7	Valor Z	-11.79570239
8	Valor P	2.0538E-32

Figura 29. Pasos para aplicar la prueba de una proporción para una cola hacia la izquierda, utilizando abordaje manual y algunas funciones. A la izquierda se presentan las fórmulas; a la derecha los valores resultantes.

En Microsoft® Excel y R

Para ejemplificar la prueba de una proporción para dos colas, se utilizará la siguiente situación hipotética: Se realiza una encuesta para determinar qué porcentaje de productores en un municipio están aplicando obras de conservación de suelo, estudios anteriores afirman que el 50% de ellos lo están haciendo. De 46 encuestas aplicadas, se determina que 22 respondieron positivamente, a la opción que refleja la aplicación de obras de conservación. Se pretende comparar la proporción muestreada con la descrita por los estudios anteriores.

Antes de iniciar, evaluaremos si los datos cumplen los dos requisitos para aproximar la distribución binomial, utilizando la distribución normal:

$$n \times p \geq 10 \text{ y } n \times (1 - p) \geq 10$$

$$46 \times 0.5 = 23 \text{ y } 46 \times (1 - 0.5) = 23$$

Ambos valores resultantes (23 y 23) son mayor que 10, por lo que se cumplen los criterios para utilizar la distribución normal como aproximación.

Como la dirección de la prueba es hacia los dos lados (dos colas), calcularemos el valor de “p” de la cola izquierda con la corrección (+0.5) y el valor de “p” de la cola derecha con la corrección (-0.5), luego seleccionaremos el menor de los dos valores y lo multiplicamos por 2.

Para iniciar la prueba calculamos la proporción de la muestra, dividiendo el número de personas encuestadas que declararon estar realizando obras de conservación de suelo entre el total de encuestados o sea $\frac{22}{46} = 0.478260869$. Sin embargo, para la prueba utilizaremos la proporción con la corrección hacia la izquierda (+0.5) y hacia la derecha (-0.5), o sea $\frac{22 + 0.5}{46} = \frac{22.5}{46} = 0.489130435$ (celda B7) y $\frac{22 - 0.5}{46} = \frac{21.5}{46} = 0.467391304$ (celda C7) (Paso 1) (Figura 30).

Seguidamente calculamos el valor del estadístico para la cola izquierda. La fórmula de Z en MS Excel sería: $=(B7-B6)/\text{RAIZ}((B6*(1-B6))/B3)$ (en B3 se encuentra el número de observaciones y en B6 la proporción teórica), lo cual obtenemos como resultado -0.147441956 (celda B8); para la cola derecha la fórmula de Z en MS Excel sería: $=(C7-C6)/\text{RAIZ}((C6*(1-C6))/C3)$, lo cual obtenemos como resultado -0.442325868 (celda C8) (Paso 2).

Aplicaciones de Estadística Básica

El abordaje manual para calcular los valores de “p” para ambas colas es:

Cola izquierda: $=\text{DISTR.NORM.ESTAND.N}(B8, \text{VERDADERO}) = 0.441391596$ (celda B9) (Paso 3).

Cola derecha: $=1 - \text{DISTR.NORM.ESTAND.N}(C8, \text{VERDADERO}) = 0.670873293$ (celda C9) (Paso 3).

Para las dos colas: $=\text{MIN}(B9:C9)*2 = 0.882783191$ (Paso 4). Notemos el uso de la función “MIN()” para seleccionar automáticamente el menor valor entre los dos resultados de “p” (ubicados en B9 y C9).

Dado $\alpha = 0.05$, el valor de p es mayor que 0.05, por lo que fallamos en rechazar la hipótesis nula y concluimos que realmente la proporción de encuestados que están haciendo obras de conservación de suelo es igual al reportado por otros estudios (50%).

Paso 1					Paso 1				
1	A	B	C	D	1	A	B	C	
2	Respuestas +	=22+0.5	Cola Izquierda		2	Respuestas +		Cola Izquierda	Cola Derecha
3	Muestra (n)	46			3	Muestra (n)	22.5	46	21.5
4	Hipótesis 0	p=0.5			4	Hipótesis 0	p=0.5		
5	Hipótesis A	p≠0.5			5	Hipótesis A	p≠0.5		
6	p teórico	0.5			6	p teórico		0.5	0.5
7	p muestreado	=B2/B3			7	p muestreado	0.489130435		0.467391304
8	Valor Z	=(B7 - B6)/RAIZ((B6*(1-B6))/B3)			8	Valor Z	-0.147441956		-0.442325868
9	Valor P 1 Cola	=DISTR.NORM.ESTAND.N(B8,VERDADERO)			9	Valor P 1 Cola	0.441391596		0.670873293
10	Valor P 2 Colas	=MIN(B9:C9)*2			10	Valor P 2 Colas	0.882783191		

Figura 30. Pasos para aplicar la prueba de una proporción para dos colas, utilizando abordaje manual y algunas funciones. A la izquierda se presentan las fórmulas; a la derecha los valores resultantes.

Pruebas de dos proporciones

Estas pruebas comparan dos proporciones, generalmente proveniente de experimentos o muestreos (observaciones). La prueba de dos proporciones asume las siguientes hipótesis:

$$H_0: \hat{p}_1 = \hat{p}_2$$

$$H_1: \hat{p}_1 \neq \hat{p}_2; \text{ también llamada de dos colas.}$$

$$H_1: \hat{p}_1 > \hat{p}_2; \text{ también llamada de una cola hacia la derecha.}$$

$$H_1: \hat{p}_1 < \hat{p}_2; \text{ también llamada de una cola hacia la izquierda.}$$

Donde,

\hat{p}_1 = Proporción uno.

\hat{p}_2 = Proporción dos.

En dependencia del tipo de prueba de proporciones (una o dos colas), así se utilizan las fórmulas y funciones en MS Excel. El cuadro 5 presenta cada situación.

Para ejemplificar la prueba de dos proporciones para dos colas, se utilizará la siguiente situación hipotética: Se realizan encuestas en dos comunidades rurales para determinar si las proporciones de aceptación de una reforma a la ley ambiental es similar entre ellas. En la comunidad 1 se aplicaron 77 encuestas de las cuales 34 encuestados afirman están de acuerdo; en la comunidad 2 se aplicaron 68 encuestas, de las cuales 56 encuestados están de acuerdo. Se pretende comparar las proporciones de encuestados que están de acuerdo con la propuesta, entre ambas comunidades.

Aproximaremos esta distribución binomial (de acuerdo versus no de acuerdo) utilizando la distribución normal. Para ello, nuestros datos tienen que llenar dos requisitos:

$$n \times \bar{p} \geq 10 \text{ y } n \times (1 - \bar{p}) \geq 10$$

Donde,

n = Número de observaciones.

\bar{p} = Proporción combinada.

La proporción combinada se calcula con la siguiente fórmula:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Donde,

x_1 = Número de eventos con la condición requerida en la primera muestra, en celda B2 (Figura 31).

x_2 = Número de eventos con la condición requerida en la segunda muestra, en celda C2.

n_1 = Número total de observaciones en la primera muestra, en celda B3.

n_2 = Número total de observaciones en la segunda muestra, en celda C3.

Así es que primero calcularemos \bar{p} y luego probaremos si los datos llenan los dos requisitos. La fórmula de \bar{p} estaría expresada en MS Excel como: =(B2+C2)/(B3+C3), lo cual es igual a 0.620689655 (celda B8) (Paso 1).

Aplicaciones de Estadística Básica

Con \bar{p} calculado, podemos verificar los requisitos para ambas muestras por separado.

Para la muestra 1: $77 \times 0.62 = 47.7$ y $77 \times (1-0.62) = 29.3$

Para la muestra 2: $68 \times 0.62 = 42.2$ y $68 \times (1-0.62) = 25.8$

Los valores resultantes de las operaciones para ambas muestras resultaron ser mayores que 10, por lo que se cumplen los criterios para utilizar la distribución normal como aproximación.

Como el objetivo es de comparación, se asume $H_0: \hat{p}_1 = \hat{p}_2$ y $H_1: \hat{p}_1 \neq \hat{p}_2$; por lo que la prueba es de dos colas. Para esto tendremos que determinar el valor de la probabilidad “p” para la cola izquierda y derecha; luego seleccionar el menor valor y multiplicarlo por el número 2. Debemos tener en cuenta la “Corrección de Continuidad”, que para la prueba de dos proporciones consistirá en sumarle 0.5 al valor con la característica de interés acertados (en este caso los que dijeron estar de acuerdo) en el cual la proporción es menor y restarle 0.5 al valor con la característica de interés acertados en el cual la proporción es mayor. Se recomienda poner los datos de la menor proporción primero y luego los de la mayor proporción. Las proporciones para ambas muestras son:

$$\text{Comunidad 1: } \hat{p}_1 = \frac{34}{77} = 0.441558441$$

$$\text{Comunidad 2: } \hat{p}_2 = \frac{56}{68} = 0.823529411$$

Sin embargo, en la prueba utilizaremos las proporciones con las correcciones para ambas comunidades, o sea:

$$\text{Comunidad 1: } \hat{p}_1 = \frac{34+0.5}{77} = 0.448051948 \text{ (celda B6) (Paso 2) (Figura 31)}$$

$$\text{Comunidad 2: } \hat{p}_2 = \frac{56-0.5}{68} = 0.816176470 \text{ (celda B7)}$$

Seguidamente calculamos \bar{q} con la fórmula: $\bar{q} = 1 - \bar{p}$, la cual quedaría expresada en MS Excel como=1- B8 (en la celda B8 se encuentra el valor de \bar{p}), lo que daría como resultado 0.3793103 (celda B9) (Paso 3).

A continuación calculamos el valor del estadístico Z utilizando la fórmula:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

Donde,

\hat{p}_1 = Proporción 1, en celda B6

\hat{p}_2 = Proporción 2, en celda B7

\bar{p} = Proporción combinada (p barra), en celda B8

\bar{q} = 1 Proporción combinada (q barra), en celda B9

n_1 = Número total de observaciones en la primera muestra, en celda B3

n_2 = Número total de observaciones en la segunda muestra, en celda C3

La fórmula de Z la expresaríamos en MS Excel como:

=(B6-B7)/RAIZ((B8*B9/B3)+(B8*B9/C3)), la que da como resultado -4.559066945 (celda B10) (Paso 4).

Luego determinamos el valor de la probabilidad, calculando primero el valor de “p” para una cola hacia la izquierda, luego para una cola hacia la derecha y finalmente seleccionamos el menor valor de ambos “p” para multiplicarlo por 2. El abordaje quedaría resumido en:

=DISTR.NORM.ESTAND.N(B10,VERDADERO) = 2.56907E-06 (celda B11) (Paso 5)

=1-B11 = 0.999997431 (celda B12) (Paso 6)

=MIN(B11:B12)*2 = 5.13814E-06 (Paso 7), es decir 5.13814×10^{-06}

Dado $\alpha = 0.05$, el valor de p es mucho menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que realmente hay diferencias significativas entre la proporción de personas que dijeron están de acuerdo a la reforma a la ley ambiental. En la comunidad 2 la mayoría están de acuerdo, en contraste con la opinión de los encuestados en la comunidad 1.

Aplicaciones de Estadística Básica

	A	B	C	D		A	B	C
1		Comunidad1		Comunidad2	1		Comunidad1	Comunidad2
2	Respuestas SI	=34+0.5		=56-0.5	2	Respuestas SI	34.5	55.5
3	Muestra (n)	77		68	3	Muestra (n)	77	68
4	Hipótesis 0	p1=p2			4	Hipótesis 0	p1=p2	
5	Hipótesis A	p1≠p2			5	Hipótesis A	p1≠p2	
6	Proporción 1	=B2/B3			6	Proporción 1	0.448051948	
7	Proporción 2	=C2/C3			7	Proporción 2	0.816176471	
8	p-barra	=(B2+C2)/(B3+C3)			8	p-barra	0.620689655	
9	q-barra	=1-B8			9	q-barra	0.379310345	
10	Valor Z	=(B6-B7)/RAIZ((B8*B9/B3)+(B8*B9/C3))			10	Valor Z	-4.559066945	
11	Valor P (1C-IZ)	=DISTR.NORM.ESTAND.N(B10,VERDADERO)			11	Valor P (1C-IZ)	2.56907E-06	
12	Valor P (1C-DE)	=1-B11			12	Valor P (1C-DE)	0.999997431	
13	Valor P (2C)	=MIN(B11:B12)*2			13	Valor P (2C)	5.13814E-06	
14					14			
15					15			
16					16			

Figura 31. Pasos para aplicar la prueba de dos proporciones para dos colas, utilizando abordaje manual y algunas funciones. A la izquierda se presentan las fórmulas; a la derecha los valores resultantes.

Prueba de bondad de ajuste

La prueba de bondad de ajuste es muy útil y popular para determinar si un conjunto de frecuencias observadas se ajustan a un conjunto de proporciones predefinidas. La prueba de bondad de ajuste asume las siguientes hipótesis:

H_0 : La variable aleatoria sigue la distribución conocida.

H_1 : La variable aleatoria sigue una distribución diferente.

O sea, la hipótesis nula (H_0) asume que los datos observados se ajustan a la distribución sugerida; y la hipótesis alternativa (H_1) supone que los datos observados no se ajustan a la distribución sugerida, por lo que tienen diferente distribución.

Para ejemplificar la prueba de bondad de ajuste, se utilizará la siguiente situación hipotética: Se conoce que las semillas de una planta, que produce frutos para la alimentación humana, al ser sembradas producirán cuatro condiciones de frutos: el 56% de las plantas que crezcan producirán frutos grandes y sin semillas, el 19% tendrán frutos grandes y con semillas, el 19% producirá frutos pequeños y con semillas y solamente un 6% producirá frutos pequeños y sin semillas.

El propietario de una finca de grandes dimensiones, donde se cultivan dichas plantas, quiere determinar si las plantas están produciendo frutos con las características y proporciones que le había ofrecido el vendedor de las semillas. Entonces, se realiza un

muestreo donde de forma aleatoria se seleccionan 5667 plantas, se miden los frutos y se revisa la presencia o ausencia de semillas, de tal forma que de las 5667 plantas: 3200 produjeron frutos grandes y sin semillas, 1057 frutos grandes y con semillas, 1072 frutos pequeños y con semillas, y 338 frutos pequeños y sin semillas. Se precisa saber si las frecuencias encontradas en el muestreo se ajustan a las proporciones dadas teóricamente.

La ecuación que se utilizará para desarrollar el análisis será la siguiente:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Donde,

X^2 = Estadístico Chi-Cuadrado o Ji al Cuadrado.

O = Valores observados.

E = Valores esperados.

Para el abordaje a mano, a fin de implementar la fórmula en MS Excel, en primer lugar, transformamos los porcentajes en proporciones, dividiendo cada uno entre 100 (Paso 1) (Figura 32). Teniendo arreglado los datos por características, datos observados y proporciones, procedemos a calcular los valores esperados a partir de la multiplicación de los valores de cada proporción por el total de observaciones (5667) (Paso 2); luego aplicamos la fórmula del estadístico Chi-Cuadrado, la cual para el primer valor quedaría expresada en MS Excel como $=((B2-D2)^2)/D2$, donde en la celda B2 se encuentra el primer valor observado y en la celda D2 se encuentra el primer valor esperado, la fórmula la aplicamos con al resto de los valores con la opción de relleno (Paso 3); consecutivamente sumamos los resultados mediante la función "SUMA()" y así obtenemos el valor del estadístico Chi-Cuadrado, el cual es 0.615261906 (Paso 4). Finalmente calculamos la probabilidad utilizando la función $=PRUEBA.CHICUAD(B2:B5, D2:D5)$ (Paso 5), donde B2:B5 es el rango de los datos observados y D2:D5 es el rango de los datos esperados, el resultado de la función es 0.892929793.

Dado $\alpha = 0.05$, el valor de p es mucho mayor que 0.05, por lo que fallamos en rechazar la hipótesis nula y concluimos que las frecuencias observadas se ajustan a las proporciones teóricas que denotaban las características de los frutos en las plantas en cuestión.

Paso 2

G15

X

✓

f_x

A

B

C

D

E

1

2

3

4

5

6

7

8

9

Características

Observado

Proporciones

Esperados

(O-E)²/E

Grand-SinSemi

3200

0.56

=C2*5667

=(B2-D2)^2/D2

Grand-ConSemi

1057

0.19

=C3*5667

=(B3-D3)^2/D3

Peque-ConSemi

1072

0.19

=C4*5667

=(B4-D4)^2/D4

Pequeñ-SinSemi

338

0.06

=C5*5667

=(B5-D5)^2/D5

SumaObser

=SUMA(B2:B5)

Chi cuadrado

=SUMA(E2:E5)

Valor P

=PRUEBA.CHICUAD(B2:B5, D2:D5)

Paso 1

Paso 3

Paso 4

G15

X

✓

f_x

A

B

C

D

E

1

2

3

4

5

6

7

8

9

Características

Observado

Proporciones

Esperados

(O-E)²/E

Grand-SinSemi

3200

0.6

3173.52

0.220950364

Grand-ConSemi

1057

0.2

1076.73

0.36153251

Peque-ConSemi

1072

0.2

1076.73

0.020778561

Pequeñ-SinSemi

338

0.1

340.02

0.012000471

SumaObser

5667

Chi cuadrado

0.615261906

Valor P

0.892929793

Paso 5

Figura 32. Pasos para aplicar la prueba bondad de ajuste, utilizando abordaje manual y algunas funciones. A la izquierda se presentan las fórmulas; a la derecha, los valores resultantes.

Pruebas de independencia (tablas de contingencia)

Las pruebas de independencia tienen como objetivo determinar la relación entre variables categóricas utilizando datos de frecuencias. En este escrito, se hará referencia a dos tipos de pruebas de independencia, las de 2 x 2 y las de R x C. Los nombres denotan el número de niveles de factores que se contrastan.

En las pruebas de 2 x 2 se comparan cuatro niveles, dos de un factor y dos del otro, de las cuales dos están ubicadas en las filas y dos en las columnas (2 x 2); por otro lado, en las pruebas de R x C se comparan “n” número niveles de un factor y “n” del otro, de las cuales los niveles de un factor están ubicadas en las filas y los niveles del otro en las columnas (R x C). Las letras R y C representan las palabras en lengua Inglesa “Rows” y “Columns” o filas y columnas. Ambos tipos de tablas utilizan la prueba Chi-Cuadrado (X^2) y similar procedimiento.

Tablas de contingencia 2 x 2

La prueba compara dos factores (o variables) con dos niveles cada uno, asumiendo las siguientes hipótesis:

H_0 : Los dos factores (o variables) son independientes.

H_1 : Los dos factores (o variables) son dependientes.

O sea, la hipótesis nula (H_0) asume que los dos factores o variables no tienen ninguna relación; y la hipótesis alternativa (H_1) supone que los dos factores o variables si tienen relación.

Para ejemplificar el uso de las tablas de contingencia 2 x 2, se utilizará la siguiente situación hipotética: Se realizan encuestas a dos comunidades sobre el estar o no estar de acuerdo con el establecimiento de una empresa minera en sus territorios. Sin embargo, existe la sospecha de que, por alguna razón intrínseca, el responder “SÍ” o “NO” depende de las comunidades. Para ello se aplicaron 77 encuestas en la comunidad 1 y 68 en la comunidad 2. Los resultados del número de personas que respondió “SÍ” o “NO” se muestra en figura 33. Se pretende determinar si hay una relación entre las respuestas y las comunidades.

En este sentido tenemos dos factores, el primero es “La Respuesta”, la cual tiene dos niveles: “SÍ” o “NO”; el segundo es “La Comunidad”, el cual tiene dos niveles: “Comunidad 1” y “Comunidad 2”. De esto, se estructura la tabla de contingencia de 2 x 2.

Lo primero que debemos hacer es estructurar la tabla con sus frecuencias observadas, estableciendo las comunidades en las columnas y las respuestas en las filas (o viceversa) (Paso 1) (Figura 33). Seguidamente se calculan los totales, tanto de las filas como de las columnas con la función “SUMA()” (Totales_Filas, Totales Columnas, respectivamente); además, calculamos el total general, sumando los resultados de los totales de las filas o de los totales de las columnas (Paso 2).

El siguiente paso es el calcular los valores esperados utilizando la ecuación:

$$Esperados = \frac{Totales_Filas \times Totales_Columnas}{Total\ General}$$

La ecuación quedaría expresada en MS Excel para el primer valor como =D3*77/145 (celda B8) lo que es igual a 47.79310345, así sucesivamente se calculan los otros valores esperados (Paso 3). Ahora estamos preparados para calcular el estadístico Chi-Cuadrado con la siguiente fórmula:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Donde,

X^2 = Estadístico Chi-Cuadrado.

O= Valores observados.

E= Valores esperados.

Aplicaciones de Estadística Básica

Sin embargo, dado que los grados de libertad para este tipo de tablas es 1, se tiene que aplicar una corrección llamada “Corrección de Yates”, esta consiste en sustraer 0.5 al valor absoluto resultante de la resta entre los valores observados, menos los esperados. Entonces, la nueva fórmula quedaría expresada como:

$$X^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

Si se desea verificar que los grados de libertad es solamente 1, se puede aplicar la siguiente fórmula para determinarlo:

$$gl = (\# \text{ filas} - 1)(\# \text{ columnas} - 1)$$

Donde,

gl = Grados de libertad.

filas = Número de filas (2).

columnas = Número de columnas (2).

Desarrollando la ecuación los grados de libertad serían: $gl = (2 - 1)(2 - 1) = 1$.

Para el primer valor, la fórmula quedaría expresada en MS Excel como: $= (ABS(B3 - B8) - 0.5)^2 / B8$. El valor observado se encuentra en la celda B3 y el valor esperado en la celda B8; la función “ABS()” devuelve el valor absoluto. El resultado es 3.697324227, así se aplica la operación para los valores restantes (Paso 4).

A continuación sumamos los cuatro valores resultantes, ubicados en el rango (B12:C13) y obtenemos 20.78509141, que sería el valor del estadístico Chi-Cuadrado (celda B15) (Paso 5). Finalmente calculamos el valor de la probabilidad (p) utilizando la función “DISTR.CHICUAD()”, el resultado de la función se la restaremos a 1 y dentro de la función le indicaremos al programa el valor del estadístico Chi-Cuadrado, los grados de libertad y el argumento “VERDADERO” que devuelve la función de distribución acumulativa. La función completa quedaría expresada como: $= 1 - DISTR.CHICUAD(B15, 1, VERDADERO)$ (Paso 6). El resultado de la aplicación de la función es 5.13814E-06, es decir 5.13814×10^{-06}).

Dado $\alpha = 0.05$, el valor de p es mucho menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que los factores son dependientes (tienen relación), por lo tanto el responder “SÍ” o “NO” esta significativamente en dependencia del tipo de comunidad (1 o 2).

Paso 1				Paso 2			
OBSERVADOS	Comunidad1	Comunidad2	Totales_Filas	OBSERVADOS	Comunidad1	Comunidad2	Totales_Filas
Si	34	56	=SUMA(B3:C3)	Si	34	56	90
No	43	12	=SUMA(B4:C4)	No	43	12	55
Totales_Columnas	=SUMA(B3:B4)	=SUMA(C3:C4)	=SUMA(B5:C5)	Totales_Columnas	77	68	145
ESPERADOS	Comunidad1	Comunidad2		ESPERADOS	Comunidad1	Comunidad2	
Si	=D3*77/145	=D3*68/145		Si	47.79310345	42.20689655	
No	=D4*77/145	=D4*68/145		No	29.20689655	25.79310345	
$(O-E -0.5)^2/E$				$(O-E -0.5)^2/E$			
Si	=ABS(B3-B8)-0.5)^2/B8	=ABS(C3-C8)-0.5)^2/C8		Si	3.697324227	4.186675963	
No	=ABS(B4-B9)-0.5)^2/B9	=ABS(C4-C9)-0.5)^2/C9		No	6.050166918	6.850924304	
Chi cuadrado	=SUMA(B12:C13)			Chi cuadrado	20.78509141		
Valor p	=1-DISTR.CHICUAD(B15,1,VERDADERO)			Valor p	5.13814E-06		
Paso 3				Paso 4			
Paso 5				Paso 6			

Figura 33. Pasos para aplicar la prueba de Chi-Cuadrado en tabla de contingencia 2 x 2 utilizando abordaje manual y algunas funciones. A la izquierda se presentan las fórmulas; a la derecha los valores resultantes.

Tablas de contingencia R x C

La prueba compara dos factores (o variables) con más de dos niveles cada uno, asumiendo las siguientes hipótesis:

H_0 : Los factores (o variables) son independientes.

H_1 : Los factores (o variables) son dependientes.

O sea, la hipótesis nula (H_0) asume que los factores o variables no tienen ninguna relación; y la hipótesis alternativa (H_1) supone que los factores o variables si tienen relación.

Para ejemplificar el uso de las tablas de contingencia R x C, se utilizará la siguiente situación hipotética: Se realizan encuestas a tres comunidades sobre el estar o no estar de acuerdo, con el establecimiento de una empresa minera en sus territorios. Sin embargo, existe la sospecha de que, por alguna razón intrínseca, el responder “Sí”, “NO” o “Indiferente” (o sea que se abstenga a opinar) depende de las comunidades. Para ello se aplicaron 83 encuestas en la comunidad 1; 76 en la comunidad 2 y 60 en la comunidad 3. Los resultados del número de personas que respondió “Sí”, “NO” o “Indiferente” se muestran en la figura 34. Se pretende determinar si hay una relación entre las respuestas y las comunidades.

Aplicaciones de Estadística Básica

En este sentido tenemos dos factores, el primero es “La Respuesta”, la cual tiene tres niveles: “SÍ”, “NO” o “Indiferente”; el segundo factor es “La Comunidad”, la cual tiene tres niveles: “Comunidad 1”, “Comunidad 2” y “Comunidad 3”. De esto, se estructura la tabla de contingencia de R x C.

Lo primero que debemos hacer es estructurar la tabla con sus frecuencias observadas, estableciendo las comunidades en las columnas y las respuestas en las filas (o viceversa) (Paso 1) (Figura 34). Seguidamente se calculan los totales, tanto de las filas como de las columnas, con la función “SUMA()” (Totales_Filas, Totales Columnas, respectivamente); además, calculamos el total general sumando los resultados de los totales de las filas o de los totales de las columnas (Paso 2).

El siguiente paso es el calcular los valores esperados utilizando la ecuación:

$$Esperados = \frac{Totales_Filas \times Totales_Columnas}{Total\ General}$$

La ecuación quedaría expresada en MS Excel para el primer valor como =E3*83/219 (celda B9) lo que es igual a 38.65753425, así sucesivamente se calculan los otros valores esperados (Paso 3). Ahora estamos preparados para calcular el estadístico Chi-Cuadrado con la siguiente fórmula:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Donde,

X^2 = Estadístico Chi-Cuadrado.

O= Valores observados.

E= Valores esperados.

En MS Excel la fórmula quedaría expresada para el primer valor como: =(B3-B9)^2/B9, donde en la celda B3 se encuentra el primer valor observado y el B9 se encuentra el primer valor esperado, dando como resultado 0.561148704. La fórmula la aplicamos a los restantes valores (Paso 4), cuyos resultados se ubican en el rango (B14:D16) que posteriormente los sumamos y el número resultante sería el estadístico de Chi-Cuadrado, en este caso el valor es 45.09527328 (celda B18) (Paso5).

Consecutivamente calculamos el valor de la probabilidad (p) utilizando la función “=PRUEBA.CHICUAD()”, escribiendo como argumento primero el rango de los datos observados y como segundo argumento el rango de datos esperados, o sea: “=PRUEBA.CHICUAD()”(B3:D5, B9:D11) (Paso 6). El resultado de la aplicación de la función es

3.79868E-09, es decir 3.79868×10^{-09} .

Dado $\alpha = 0.05$, el valor de p es mucho menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que los factores son dependientes (tienen relación), por lo tanto el responder “SÍ”, “NO” o “Indiferente” está significativamente en dependencia del tipo de comunidad (1, 2 o 3).

Paso 1					Paso 2						
A	B	C	D	E	A	B	C	D	E		
1	OBSERVADOS				1	OBSERVADOS					
2		Comunidad1	Comunidad2	Comunidad3	Totales_Filas	2		Comunidad1	Comunidad2	Comunidad3	Totales_Filas
3	Si	34	56	12	=SUMA(B3:D3)	3	Si	34	56	12	102
4	No	43	12	38	=SUMA(B4:D4)	4	No	43	12	38	93
5	Indiferente	6	8	10	=SUMA(B5:D5)	5	Indiferente	6	8	10	24
6	Totales_Columnas	=SUMA(B3:B5)	=SUMA(C3:C5)	=SUMA(D3:D5)	=SUMA(B6:D6)	6	Totales_Columnas	83	76	60	219
7	ESPERADOS					7	ESPERADOS				
8		Comunidad1	Comunidad2	Comunidad3		8		Comunidad1	Comunidad2	Comunidad3	
9	Si	=E3*83/219	=E3*76/219	=E3*60/219		9	Si	38.65753425	35.39726027	27.94520548	
10	No	=E4*83/219	=E4*76/219	=E4*60/219		10	No	35.24657534	32.27397726	25.47945205	
11	Indiferente	=E5*83/219	=E5*76/219	=E5*60/219		11	Indiferente	9.095890411	8.328767123	6.575342466	
12	(O-E) ² /E					12	(O-E) ² /E				
13		Comunidad1	Comunidad2	Comunidad3		13		Comunidad1	Comunidad2	Comunidad3	
14	Si	=(B3-B9)^2/B9	=(C3-C9)^2/C9	=(D3-D9)^2/D9		14	Si	0.561148704	11.99168752	9.098146656	
15	No	=(B4-B10)^2/B10	=(C4-C10)^2/C10	=(D4-D10)^2/D10		15	No	1.705572622	12.73577226	6.152570334	
16	Indiferente	=(B5-B11)^2/B11	=(C5-C11)^2/C11	=(D5-D11)^2/D11		16	Indiferente	1.053721736	0.01297765	1.783675799	
17						17					
18	Chi cuadrado	=SUMA(B14:D16)				18	Chi cuadrado	45.09527328			
19	Valor p	=PRUEBA.CHICUAD(B3:D5,B9:D11)				19	Valor p	3.79868E-09			
20						20					
21						21					

Figura 34. Pasos para aplicar la prueba de Chi-Cuadrado en tabla de contingencia R x C utilizando abordaje manual y algunas funciones. A la izquierda se presentan las fórmulas; a la derecha los valores resultantes.

Comparación de medias

En términos de comparación de media se describirán dos tipos de pruebas, un conjunto de pruebas para comparar las medias de dos grupos y un conjunto de pruebas para comparar las medias de más de dos grupos. Específicamente abordaremos la “prueba T” y sus varios tipos, y el “análisis de varianza” y algunos de sus tipos.

Antes de iniciar a explicar la prueba T, vamos a describir la prueba F. La prueba F será útil para verificar si dos conjuntos de datos cumplen o no con el supuesto de igualdad de varianza. MS Excel cuenta con una opción dentro de las herramientas de análisis de datos, con la que automáticamente se puede aplicar la prueba. Esta asume las hipótesis:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Aplicaciones de Estadística Básica

Donde,

σ_1^2 = Varianza del grupo 1.

σ_2^2 = Varianza del grupo 2.

Para ejemplificar su aplicación, utilizaremos unos datos de humedad del suelo (%) tomados en 10 puntos aleatorios en dos sitios (10 en el Sitio 1 y 10 en el Sitio 2). Para aplicar la prueba F a estos datos, nos dirigimos a la opción “Análisis de datos” la cual se encuentra en “DATOS” (Paso 1) (Figura 35) y seleccionamos la opción “Prueba F para varianzas de dos muestras” (Paso 2), aceptamos y nos aparecerá un nuevo cuadro de diálogo, en este hacemos clic en la flecha roja utilizada para definir el rango del primer grupo de datos (Paso 3), el cuadro de diálogo se minimizará y podemos seleccionar los datos (Paso 4), luego maximizamos el cuadro de diálogo (Paso 5). A continuación, hacemos clic en la flecha roja utilizada para definir el rango del segundo grupo de datos (Paso 6) y procedemos como en los pasos 4 y 5.

Luego ponemos check en “Rótulos” (Paso 7) pues las columnas tienen encabezados, el “Alfa” lo dejamos por defecto (o se cambia según convenga). Finalmente establecemos la opción de salida llamada “Rango de salida” (preferencia personal) (Paso 8) y le indicamos al programa la celda en donde deseamos que se plasme el cuadro de resultados, haciendo clic en la flecha roja que está al lado derecho (Paso 9), seleccionamos la celda (D1 en el ejemplo), maximizamos el cuadro de diálogo y hacemos clic en “Aceptar” (Paso 10).

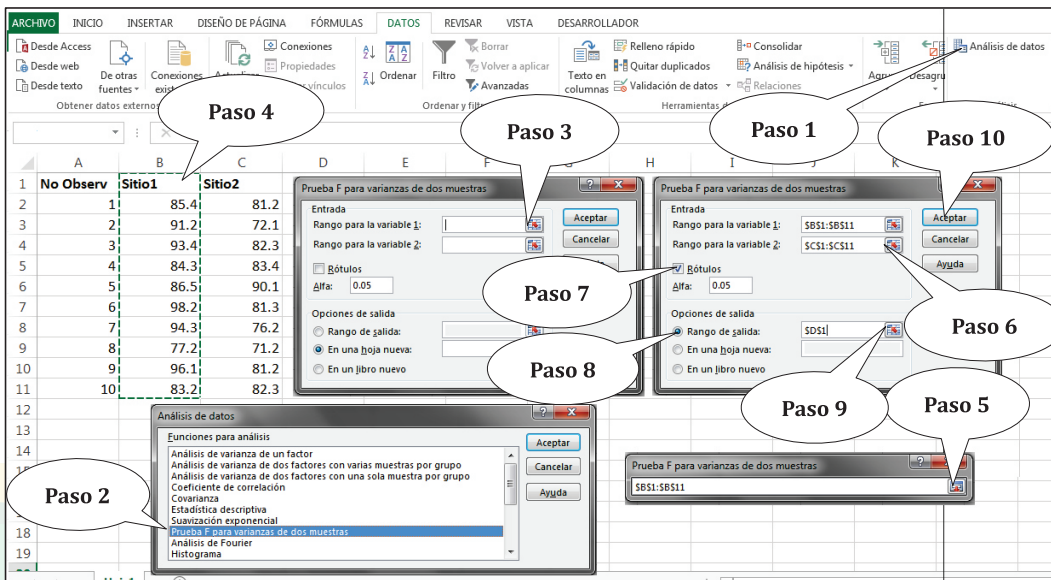


Figura 35. Ilustración de los pasos para aplicar prueba F. Notemos los dos conjuntos de datos en las columnas B y C.

En la figura 36 se muestra la tabla de resultados, con información sobre la prueba F. Si nos dirigimos al valor de “p”, representado por $P(F \leq f)$, notamos que el valor es 0.305 para una cola; este lo multiplicamos por 2 para dos colas y así el valor para nuestra prueba es 0.61011135. Dado $\alpha = 0.05$, el valor de “p” es mucho mayor por lo que fallamos en rechazar H_0 y, por lo tanto, concluimos que hay igualdad de varianzas.

Prueba F para varianzas de dos muestras		
	Sitio1	Sitio2
Media	88.98	80.13
Varianza	44.6128889	31.4267778
Observaciones	10	10
Grados de libertad	9	9
F	1.41958203	
$P(F \leq f)$ una cola	0.30505567	
Valor crítico para F (una cola)	3.1788931	

Figura 36. Resultados de la prueba F, presenta estadística descriptiva, el estadístico F, el valor de “p” y el valor crítico para F.

Prueba T para una muestra

La prueba T para una muestra compara la media de un conjunto de valores con un valor teórico preestablecido. Esta prueba asume las siguientes hipótesis:

$$H_0: \bar{x} = \mu$$

$$H_1: \bar{x} \neq \mu; \text{ también llamada de dos colas.}$$

$$H_1: \bar{x} > \mu; \text{ también llamada de una cola hacia la derecha.}$$

$$H_1: \bar{x} < \mu; \text{ también llamada de una cola hacia la izquierda.}$$

Donde,

\bar{x} = Media del conjunto de datos

μ = Dato teórico

En dependencia del tipo de prueba T (una o dos colas), así se utilizan las fórmulas y funciones en MS Excel. En el cuadro 6 se presenta cada situación.

Aplicaciones de Estadística Básica

Cuadro 6. Tipo colas, fórmulas y funciones asociadas para lograr los cálculos.

COMPARACIONES	FÓRMULA
Una cola - hacia la izquierda	=DISTR.T.N(t,gl,VERDADERO)
Una cola - hacia la derecha	=DISTR.T.CD(t,gl)
Dos colas	=DISTR.T.2C(t,gl) Para t<0 devuelve “#¡NUM!”, usar valor absoluto.

Si el argumento acumulado es VERDADERO, DISTR.T.N devuelve la función de distribución acumulativa; si es FALSO, devuelve la función de densidad de probabilidad.

t= Estadístico de la prueba T.

Para ejemplificar la prueba T para una cola hacia la izquierda, se utilizará la siguiente situación hipotética: Los niveles de oxígeno disuelto (OD) en el agua no pueden ser menores de 3 ppm (partes por millón), de serlo, toda la fauna acuática estaría en peligro. A lo largo de una fuente de agua se realizaron siete muestreos y se determinó el OD a cada una, de tal forma que se pretende comparar la media de los datos observados con el valor de referencia (3 ppm) para determinar si los niveles de OD realmente son menores a ese valor.

El cálculo del estadístico lo realizamos aplicando la fórmula:

$$t = \frac{\bar{x} - \mu}{EE}$$

Donde,

t= El estadístico t.

\bar{x} = La media del conjunto de datos.

μ = El valor de referencia.

EE= El error estándar.

Sin embargo, antes de aplicar la formula tendríamos que calcular el valor de EE, esto requiere el uso de una fórmula en MS Excel, ya que no contamos con una función que genere el valor automáticamente:

$$EE = \frac{DE}{\sqrt{n}}$$

Donde,

EE= El error estándar.

DE= La desviación estándar.

n= Número de observaciones (muestras).

Antes de aplicar ambas fórmulas, calcularemos los datos descriptivos necesarios. En el caso del número de observaciones (n) lo calculamos con la función “CONTAR()” (valor calculado en la celda E1) (Figura 37), la media (\bar{x}) la calculamos con la función “PROMEDIO()” (en la celda E2) y la desviación estándar (DE) con “DESVEST.M()” (en la celda E3) (Paso 1). El error estándar (EE) lo calculamos entonces con la fórmula (expresada en MS Excel): $=E3/RAIZ(E1)$, cuyo resultado es 0.243416722 (en la celda E4) (Paso 2).

Adicionalmente calculamos los grados de libertad restándole 1 al número de observaciones, o sea $=E1-1$ (en la celda E6) (Paso 3). Y a continuación calculamos el estadístico t , mediante la fórmula expresada en MS Excel $=(E2-E5)/E4$ (en E5 se encuentra el valor de referencia μ), el resultado es -2.523597017 (en la celda E7) (Paso 4).

Finalmente obtenemos el valor de la probabilidad “ p ” utilizando la función de cola hacia la izquierda (debido a que $H_1: \bar{x} < \mu$): “DISTR.T.N()”, la cual quedaría expresada como: $=DISTR.T.N(E7,E6,VERDADERO)$, cuyo resultado es 0.022533807 (Paso 5). En R se reporta el valor de 0.04507 para las tres opciones de cola, siendo el equivalente a la opción de dos colas en MS Excel.

Dado $\alpha = 0.05$, el valor de p es menor que 0.05, por lo que se rechaza la hipótesis nula y se concluye que realmente el oxígeno disuelto en el cuerpo de agua presenta valores significativamente menores a 3 ppm.

El abordaje para las pruebas de dos colas y de una cola hacia la derecha, es similar al descrito en el ejemplo anterior y las diferencias radican en el uso de la fórmula para el cálculo del valor de “ p ”, dichas variaciones se muestran en el cuadro 6.

	A	B	C	D	E		A	B	C	D	E
1	Muestras	OD	n		=CONTAR(B2:B8)	Paso 1	1	Muestras	OD	n	7
2	1	2.2	Media		=PROMEDIO(B2:B8)		2	1	2.2	Media	2.385714286
3	2	3.2	DE		=DESVEST.M(B2:B8)	Paso 2	2	3.2	DE	0.644020112	
4	3	2.1	EE		=E3/RAIZ(E1)		3	2.1	EE	0.243416722	
5	4	2.3	Valor Refer	3			5	4	2.3	Valor Refer	3
6	5	3.1	gl		=E1-1	Paso 3	6	5	3.1	gl	6
7	6	2.5	t		=(E2-E5)/E4		7	6	2.5	t	-2.523597017
8	7	1.3	p		=DISTR.T.N(E7,E6,VERDADERO)	Paso 4	8	7	1.3	p	0.022533807
						Paso 5					

Figura 37. Pasos para aplicar una prueba T para una muestra y de una cola hacia la izquierda. A la izquierda se presentan las fórmulas; a la derecha los valores resultantes.

Aplicaciones de Estadística Básica

Prueba T para dos muestras independientes

Lo básico que se necesita saber de la prueba T para dos muestras independientes, es que compara dos conjuntos (grupos) de datos de una misma variable y genera un valor de significancia, que nos sirve para decidir si los conjuntos de datos son semejantes o significativamente diferentes.

La prueba T para dos muestras independientes asume las siguientes hipótesis:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$; también llamada de dos colas.

$H_1: \mu_1 > \mu_2$; también llamada de una cola hacia la derecha.

$H_1: \mu_1 < \mu_2$; también llamada de una cola hacia la izquierda.

Donde,

μ_1 = Media del conjunto de datos 1.

μ_2 = Media del conjunto de datos 2.

Por ejemplo, si se midiera la humedad del suelo en 10 puntos aleatorios en dos sitios, entonces tuviéramos dos grupos de datos, los datos del sitio 1 y los datos del sitio 2 de una sola variable (humedad del suelo en %) (Figura 38). Entonces, nos interesa determinar si hay diferencias significativas entre los dos conjuntos.

Para aplicar la prueba seleccionamos la herramienta de análisis de datos (Paso 1) (Figura 38) y de ella seleccionamos “Prueba T para dos muestras suponiendo varianzas iguales” (Paso 2) (Para este ejemplo ya hemos confirmado la igualdad de varianza). En el siguiente cuadro de diálogo le indicaremos al programa el rango de los datos del primero (Paso 3) y del segundo grupo (Paso 4), ponemos check en “Rótulos” (si las columnas tiene encabezado) (Paso 5) y “Rango de salida” (Paso 6) definiendo la celda donde se desplegará la tabla de resultados (para este ejemplo, en la celda D1) (Paso 7), las otras opciones se dejan con lo que contienen por defecto (para este ejemplo) y finalmente seleccionamos “Aceptar”.

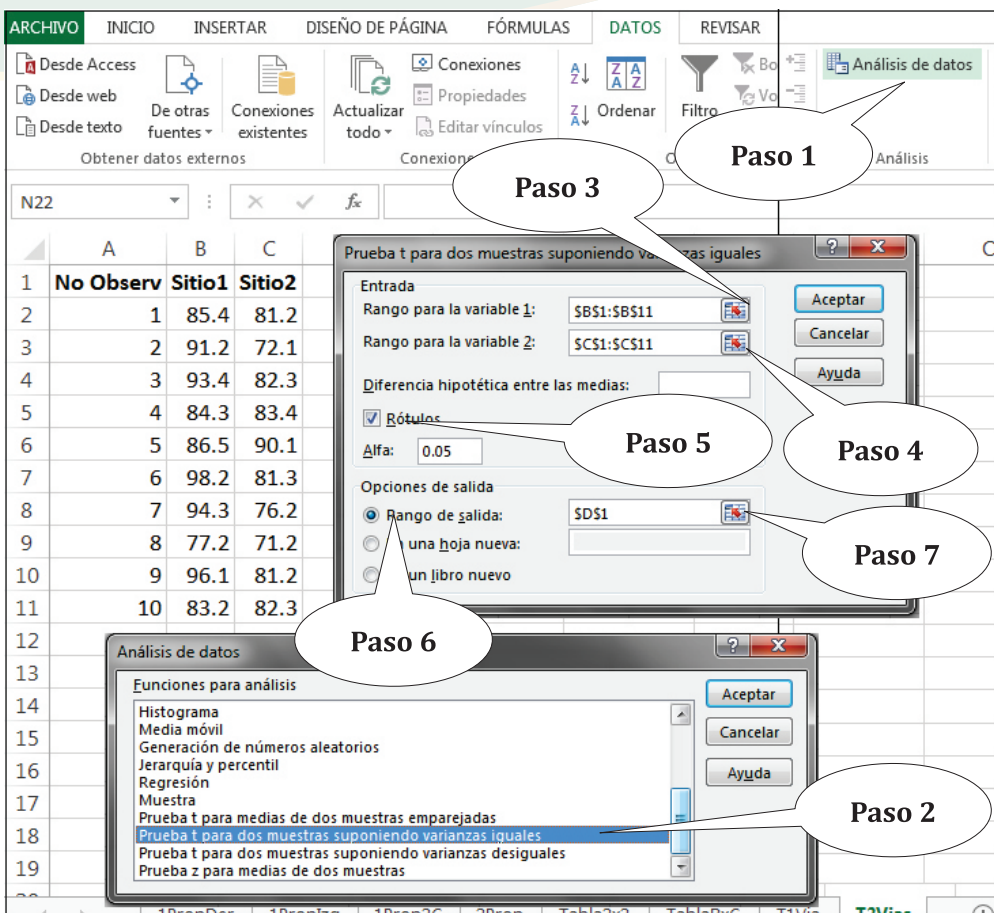


Figura 38. Pasos para aplicar un prueba T para datos independientes, en el caso particular del ejemplo, suponiendo varianzas iguales.

El cuadro de resultado en la figura 39 muestra mucha información, la cual el lector tiene la libertad de usarla o ignorarla, para objeto de este ejercicio, la atención será dirigida al “Estadístico t” y al valor de “p” denotado por “P(T<=t) dos colas”. Dado $\alpha = 0.05$, el valor de “p” (0.0049) es menor, por lo que se rechaza la hipótesis nula y concluimos que hay diferencias significativas de la humedad del suelo, entre los dos sitios muestreados, según las muestras colectadas.

D	E	F
Prueba t para dos muestras suponiendo varianzas iguales		
	<i>Sitio1</i>	<i>Sitio2</i>
Media	88.98	80.13
Varianza	44.6128889	31.4267778
Observaciones	10	10
Varianza agrupada	38.0198333	
Diferencia hipotética de las medias	0	
Grados de libertad	18	
Estadístico t	3.20939498	
P(T<=t) una cola	0.00243051	
Valor crítico de t (una cola)	1.73406361	
P(T<=t) dos colas	0.00486103	
Valor crítico de t (dos colas)	2.10092204	

Figura 39. Resultado de una prueba T para datos independientes.

Para aplicar una prueba T datos independientes, donde se asume desigualdad de varianza seguimos exactamente los mismos pasos descritos para aplicar una prueba T de datos independientes, donde se asume igualdad de varianza, con la excepción que el tipo de prueba que se selecciona en la opción de “Análisis de datos” es la que se llama “Prueba T para dos muestras suponiendo varianzas desiguales”.

La opción no paramétrica de la prueba T independiente se llama Mann-Whitney o Wilcoxon. MS Excel no ofrece una opción directa para el cálculo de esta prueba, para lo cual se tendría que instalar algún complemento especial o realizar el cálculo mediante procedimientos tediosos, no abordados en este libro. La aplicación de la prueba de Mann-Whitney o Wilcoxon se abordará en la sección de R.

Prueba T para dos muestras pareadas

La prueba T para dos muestras pareadas, compara dos conjuntos de datos (grupos) que son dependientes entre ellos, pues contienen medidas en el tiempo (generalmente) que se han tomado a los mismos objetos. Por ejemplo, cuando se toma un dato de salinidad en diferentes sitios, dentro de un manglar y tiempo después se vuelven a tomar datos en exactamente los mismos sitios para comparar las medias entre los dos conjuntos de datos. Estos dos conjuntos (momento 1 y momento 2) están pareados y ligados por los sitios y las diferencias se exploran en el tiempo.

La prueba t para dos muestras pareadas asume las siguientes hipótesis:

$$H_0: \mu d = 0$$

$H_1: \mu d \neq 0$; también llamada de dos colas.

$H_1: \mu d > 0$; también llamada de una cola hacia la derecha.

$H_1: \mu d < 0$; también llamada de una cola hacia la izquierda.

Donde,

μd = La media de las diferencias.

d = Diferencia entre los pares de medidas.

Para ejemplificar, asumiremos la medición del peso (lb) de un grupo de 10 venados antes y después de haber suministrado una dosis de desparasitantes, para parásitos internos en un plan de Manejo de Fauna Silvestre. Los venados estaban codificados y uno a uno se capturó y se midió su peso inicial, tres meses después de estar suministrando el desparasitante, en los bebederos artificiales, se volvieron a capturar y pesar. Se tiene el interés de conocer si el peso de los venados (antes y después) cambió, dicho cambio podría ser atribuido (posiblemente) al suministro del desparasitante. La información compilada se presenta en la figura 40.

Para correr la prueba haremos uso de la opción de análisis de datos (Paso 1), en la cual seleccionaremos la opción “Prueba T para medias de dos muestras emparejadas” (Paso 2). Aparecerá un cuadro de diálogos en donde se asignaran los datos de “Antes” donde dice “Rango para la variable 1” (Paso 3) y los datos de “Después” donde dice “Rango para la variable 2” (Paso 4). Seguidamente ponemos check en “Rótulos” (Paso 5) si seleccionamos también los encabezados de los conjuntos de datos; mantenemos el alfa en 0.05, excepto si es necesario cambiarlo; seleccionamos la “Opción de salida”, para este ejemplo, y por preferencia del autor, seleccionamos “Rango de salida” (Paso 6) y elegimos la celda donde aparecerá la tabla de resultados (Paso 7) (celda D1).

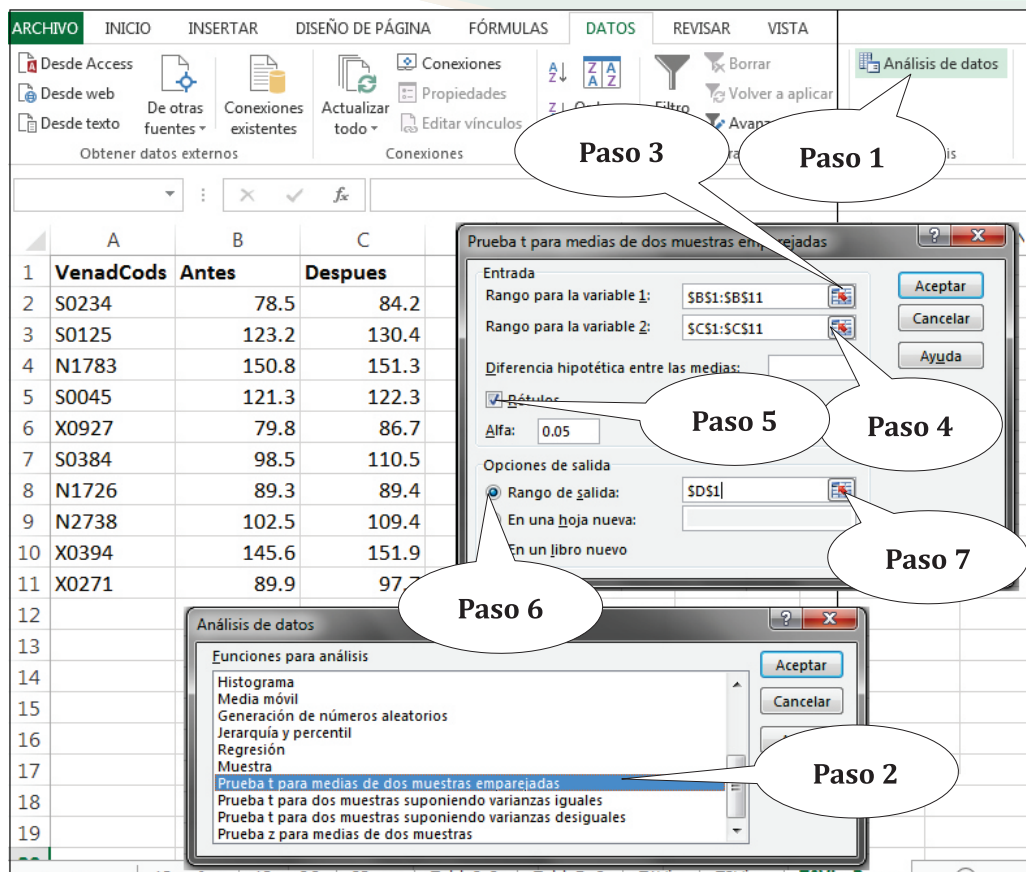


Figura 40. Pasos a seguir para aplicar una prueba T para dos muestras pareadas.

La figura 41 muestra la tabla de resultado del análisis, en esta se presentan algunas medidas descriptivas, el estadístico y los valores de la probabilidad (p) para una o dos colas. Para el ejemplo en el que se quería compara las diferencias de los pesos de los venados, en los dos momentos (dos colas) el valor de “p” es 0.0014. Dado $\alpha = 0.05$, el valor de “p” es menor que 0.05, por lo que se rechaza la hipótesis nula y se concluye que si hay diferencia entre los pesos de los venados, de un momento al otro, y conociendo los valores de las medias, notamos que los venados han ganado peso.

D	E	F
Prueba t para medias de dos muestras emparejadas		
	<i>Antes</i>	<i>Despues</i>
Media	107.94	113.38
Varianza	679.64267	633.25511
Observaciones	10	10
Coefficiente de correlación de Pearson	0.9896664	
Diferencia hipotética de las medias	0	
Grados de libertad	9	
Estadístico t	-4.5367652	
P(T<=t) una cola	0.0007062	
Valor crítico de t (una cola)	1.8331129	
P(T<=t) dos colas	0.0014124	
Valor crítico de t (dos colas)	2.2621572	

Figura 41. Tabla de resultados de la aplicación de la prueba T para dos muestras pareadas.

La opción no paramétrica de la prueba T para datos pareados (emparejados), es la prueba de Wilcoxon para datos pareados. MS Excel no ofrece una opción directa para el cálculo de esta prueba, para lo cual se tendría que instalar algún complemento especial o realizar el abordaje manual con el uso de la hoja de cálculo, no abordado en este libro. La recomendación sería el utilizar otro programa estadístico de uso libre como R.

Análisis de varianza para un factor

Lo básico que necesitamos saber del análisis de varianza (ANDEVA -ANOVA por sus siglas en inglés-) es que compara más de dos conjuntos (grupos) de datos de la misma variable y genera un valor de significancia que nos permite decidir si los conjuntos de datos son semejantes o diferentes. Utilizaremos de nuevo el ejemplo de las medidas de humedad del suelo (%), supongamos que se anexó un tercer sitio, por lo cual ahora se tienen tres grupos de datos, los datos del Sitio 1, los datos del Sitio 2 y los datos del Sitio 3 (Figura 42). El ANDEVA asume las siguientes hipótesis:

H_0 =No hay diferencias significativas entre las medias de los grupos.

H_1 =Al menos una de las medias de los grupos es diferente.

Para aplicar en ANDEVA nos dirigimos a la opción “Análisis de datos” en “DATOS” y seleccionamos la opción “Análisis de varianza de un factor” (Paso 1). Aparece un cuadro de diálogo donde tenemos que definir el rango de datos (Paso 2), el cuadro de diálogo se minimizará para poder seleccionar los datos desde la celda B1 hasta la celda D11 (Paso 3), seguidamente ampliamos el cuadro de diálogo (Paso 4) y ponemos check en “Rótulos en la primera fila” (Paso 5), luego hacemos clic en “Rango de salida” (Paso 6) y seleccionamos la celda (F1 o \$F\$1) donde se desplegará el cuadro de resultados (Paso 7) haciendo “Aceptar” al final.

	A	B	C	D
1	No Observ	Sitio1	Sitio2	Sitio3
2		1	85.4	81.2
3			91.2	72.1
4			93.4	82.3
5			84.3	83.4
6		5	86.5	90.1
7		6	98.2	81.3
8		7	94.3	76.2
9		8	77.2	71.2
10		9	96.1	81.2
11		10	83.2	82.3

Paso 3

Paso 2

Paso 5

Paso 7

Paso 6

Paso 1

Paso 4

Figura 42. Pasos para aplicar un análisis de varianza (ANDEVA) en MS Excel.

En la figura 43 se muestra el cuadro de resultados del ANDEVA. La primera tabla se llama “Resumen” y provee datos descriptivos importantes como grupos, cuenta, suma promedio y varianza. La segunda tabla se llama “Análisis de varianza” y muestra el origen de las variaciones, la suma de cuadrados, los grados de libertad, el promedio de los cuadrados, el valor de F, la probabilidad y el valor crítico para F.

En función de nuestro interés, únicamente dirigiremos la atención al valor de “p” (probabilidad), el cual es muy pequeño (0.0069) comparado con α (0.05), de tal forma que se rechaza H_0 y concluimos que al menos uno de los grupos tiene media significativamente diferente. Si observamos los promedios, podríamos deducir que el Sitio 2, es el que está provocando las diferencias, pues el valor de su media es mucho menor al de los dos sitios restantes; sin embargo, otros procedimientos son necesario seguir para llegar a conclusiones contundentes, incluyendo las pruebas de comparaciones múltiples (o a posteriori) abordadas en la sección de Estadísticas Básicas en R.

F	G	H	I	J	K	L
Análisis de varianza de un factor						
RESUMEN						
Grupos	Cuenta	Suma	Promedio	Varianza		
Sitio1	10	889.8	88.98	44.612889		
Sitio2	10	801.3	80.13	31.426778		
Sitio3	10	882.8	88.28	44.612889		
ANÁLISIS DE VARIANZA						
Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Entre grupos	484.116667	2	242.0583333	6.0187287	0.006893031	3.354130829
Dentro de los grupos	1085.873	27	40.21751852			
Total	1569.989667	29				

Figura 43. Cuadro de resultado para el análisis de varianza de un factor.

La opción no paramétrica del ANOVA se llama Kruskal-Wallis. MS Excel no ofrece una opción directa para el cálculo de esas pruebas, para lo cual se tendría que instalar algún complemento especial o realizar el cálculo mediante procedimientos tediosos, no abordados en este libro. Sin embargo, la prueba Kruskal-Wallis se aborda en la sección de Estadísticas Básicas en R.

Análisis de varianza para dos factores

Como su nombre lo indica, el análisis de varianza para dos factores (o bifactorial) es la comparación de medias entre más de dos conjuntos de datos, los cuales están agrupados en dos factores, el factor puede estar dividido en diferentes categorías o niveles. Por ejemplo, se pueden comparar las mediciones del pH del suelo basados en el factor “Uso del suelo”, y los niveles de este factor pueden ser “Uso Agrícola”, “Uso Ganadero” y “Bosque”, entonces se habrán tomado los valores del pH en cada uno de esos niveles; sin embargo, además de ello, también se puede haber tomado en cuenta si cada uno de los niveles del factor “Uso del suelo” estaban ubicados en la parte alta (cumbre), a mediana altura o en la parte baja (valle) de un gradiente de elevación (un pico montañoso o microcuenca), en cuyo caso se le anexa otro factor al que llamaremos “Elevación” con sus tres niveles: “Alto”, “Medio” y “Bajo”.

En este sentido los valores del pH del suelo se habrán tomados en los tres niveles del factor “Uso del suelo” y los tres niveles del factor “Elevación”. Por involucrar dos factores se les llama “de dos factores” (Figura 44). La comparación de la combinación de los niveles de los dos factores se le denomina “interacción”. MS Excel tiene dos opciones para el ANDEVA con dos factores, una llamada “con replicación” y otra llamada “sin replicación”.

Aplicaciones de Estadística Básica

El análisis de varianza para dos factores asume las siguientes hipótesis:

H_0 =No hay diferencias significativas entre las medias de los niveles del factor 1.

H_1 =Al menos la media de uno de los niveles dentro del factor 1 es significativamente diferente.

H_0 =No hay diferencias significativas entre las medias de los niveles del factor 2.

H_1 =Al menos la media de uno de los niveles dentro del factor 2 es significativamente diferente.

H_0 =No hay interacciones significativas entre ambos factores.

H_1 =Hay interacciones significativas entre ambos factores.

		FACTOR 1: "Uso del suelo"		
		Nivel 1: Agrícola	Nivel 2: Ganadero	Nivel 3: Bosque
FACTOR 2: "Elevación"	Nivel 1: Alto	Datos de pH	Datos de pH	Datos de pH
	Nivel 2: Medio	Datos de pH	Datos de pH	Datos de pH
	Nivel 3: Bajo	Datos de pH	Datos de pH	Datos de pH

Figura 44. Esquema para ejemplificar un diseño de muestro donde se utiliza un Análisis de Varianza de dos factores.

ANDEVA de dos factores con replicación

El análisis de varianza de dos factores con replicación, aplica cuando los niveles de un factor tienen muestras replicadas (Figura 45 A). Este ANDEVA explora las diferencias entre los niveles del factor 1, entre los niveles del factor 2 y las interacciones de los factores (niveles de un factor versus los niveles del otro factor).

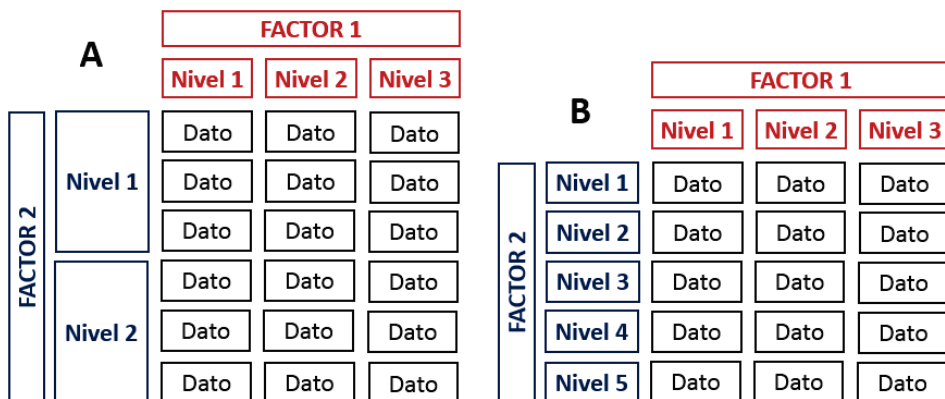


Figura 45. Ilustración de dos diseños en los que se aplican análisis de varianza de dos factores. A. Aplica ANDEVA de dos factores con replicación; B. Aplica ANDEVA de dos factores sin replicación.

Para ejemplificar, aplicaremos el análisis en la siguiente situación: Se determina la concentración de Oxígeno Disuelto (OD) (ppm) en el agua a lo largo de un río principal en una microcuenca. Los muestreos se hacen en las tres partes de la microcuenca, las cuales son: parte alta, parte media y parte baja; adicionalmente, en cada parte se selecciona dos usos del suelo por donde el río pasa, estos son el uso bosque y el uso agrícola, y en cada uno de estos usos se establecieron cinco puntos de muestreo (réplicas). Se pretende determinar si existen diferencias entre las concentraciones de OD entre las partes de las microcuencas (Factor 1), entre los usos del suelo (Factor 2) y entre sus interacciones.

El arreglo en la hoja de cálculo de MS Excel se muestra en la figura 46, las opciones para el análisis las encontraremos siguiendo la secuencia de opciones Datos>Análisis de datos y seleccionamos la opción “Análisis de varianza de dos factores con varias muestras por grupo” (Paso 1); en la opción “Rango de entrada” seleccionamos el rango de los datos, incluyendo los encabezados de columnas y nombres de filas (Paso 2); luego definimos las filas correspondientes a cada muestra replicada, en el caso de este ejemplo los datos replicados se encuentran en cinco filas por factor (Paso 3); dejaremos el mismo alfa por defecto (0.05), excepto si amerita cambio (Paso 4) e indicamos la opción de salida de los resultados, seleccionamos la opción “Rango de salida” (preferencia personal del autor) (Paso 5) y definimos la celda donde se ubicará la tabla de resultados, en este caso es la celda E1 (Paso 6).

The screenshot shows an Excel spreadsheet with data for an ANOVA analysis. The data is organized into columns A through D, with rows representing different factors and replicates. The first factor is 'Bosque' (rows 2-6) and the second factor is 'Agrícola' (rows 7-11). The columns represent different levels of the first factor: 'Alta', 'Media', and 'Baja'. The values in the cells represent the results of the analysis.

Two dialog boxes are shown, illustrating the steps to perform the analysis:

- Paso 1:** The 'Análisis de datos' dialog box is open, showing the list of functions. 'Análisis de varianza de dos factores con varias muestras por grupo' is selected.
- Paso 2:** The 'Análisis de varianza de dos factores con varias muestras por grupo' dialog box is open. The 'Rango de entrada' is set to '\$A\$1:\$D\$11', 'Fila por muestra' is set to '5', and 'Alfa' is set to '0.05'.
- Paso 3:** The 'Opciones de salida' section is visible, showing 'Rango de salida' set to '\$E\$1'.
- Paso 4:** The 'Rango de salida' is set to '\$E\$1'.
- Paso 5:** The 'Rango de salida' is set to '\$E\$1'.
- Paso 6:** The 'Rango de salida' is set to '\$E\$1'.

Figura 46. Pasos para realizar un análisis de varianza de dos factores con replicación. Notemos la estructura de tabla de datos donde se definen los niveles de cada factor. Las filas de cada muestra replicada se presentan con la palabra “Fila” para una mejor visualización, los números del 1 al 5 en rojo representan las réplicas del primer nivel del Factor 2 (Bosque) y en azul las réplicas del segundo nivel del Factor 2 (Agrícola).

Los resultados incluyen la tabla de estadísticas descriptivas por factor (resumen) y la tabla del ANDEVA (análisis de varianza), para el cual fijaremos la atención en el valor de “p” (Figura 47). La tabla muestra tres valores de “p”, uno para “Muestra” (Factor 2= Usos del suelo: Bosque y Agrícola), otro para “Columnas” (Factor 1= Partes de la microcuenca: Alta, Media y Baja) y uno para “Interacción” (Niveles del factor 1 versus niveles del factor 2).

Dado $\alpha = 0.05$, el valor de “p” es menor que 0.05 para “Muestra” y “Columnas” (0.0037 y 5.443E-6 respectivamente), por lo que rechazamos la hipótesis nulas para los dos factores y concluimos que sí hay diferencia en los valores de Oxígeno Disuelto comparado entre “Usos de suelo” y entre “Pares de la microcuenca”; el valor de “p” para “Interacción” fue mayor que 0.05 (1.0), por lo que fallamos en rechazar la hipótesis nula y descartamos algún efecto por interacciones.

E	F	G	H	I	J	K	L	M	N	O
Análisis de varianza de dos factores con varias muestras por grupo										
RESUMEN	Alta	Media	Baja	Total						
<i>Bosque</i>										
Cuenta	5	5	5	15						
Suma	23.6	15.2	14.9	53.7						
Promedio	4.72	3.04	2.98	3.58						
Varianza	0.142	0.788	0.467	1.096						
<i>Agrícola</i>										
Cuenta	5	5	5	15						
Suma	19.6	11.2	10.9	41.7						
Promedio	3.92	2.24	2.18	2.78						
Varianza	0.142	0.788	0.467	1.096						
<i>Total</i>										
Cuenta	10	10	10							
Suma	43.2	26.4	25.8							
Promedio	4.32	2.64	2.58							
ANÁLISIS DE VARIANZA										
Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F				
Muestra		4.8	1	4.8	10.307802	0.0037436	4.259677273			
Columnas		19.512	2	9.756	20.950608	5.443E-06	3.402826105			
Interacción		5.32907E-15	2	2.66454E-15	5.722E-15	1	3.402826105			
Dentro del grupo		11.176	24	0.465666667						
Total		35.488	29							

Figura 47. Resultados del análisis de varianza de dos factores con replicación. Arriba se muestran las estadísticas descriptivas, abajo la tabla del análisis de varianza.

ANDEVA de dos factores sin replicación

El análisis de varianza de dos factores sin replicación, aplica cuando los niveles de un factor no tienen muestras replicadas (Figura 45 B). Este ANDEVA explora las diferencias entre los niveles del factor 1 y entre los niveles del factor 2. Para ejemplificar, aplicaremos el análisis en la siguiente situación: Se determina la concentración de Oxígeno Disuelto (OD) (ppm) en el agua de varios ríos, que se encuentran en las tres partes de una microcuenca: parte alta, parte media y parte baja. Se pretende conocer si existen diferencias en las concentraciones de OD entre las partes de las microcuencas (Factor 1) y los ríos (Factor 2).

El arreglo en la hoja de cálculo de MS Excel se muestra en la figura 48, las opciones para el análisis se encuentran siguiendo la secuencia de opciones Datos>Análisis de datos y se selecciona la opción “Análisis de varianza de dos factores con una sola muestra por grupo” (Paso 1); en la opción “Rango de entrada” seleccionamos el rango de los datos, incluyendo los encabezados de columnas y nombres de filas (Paso 2); ponemos check

Aplicaciones de Estadística Básica

donde dice “Rótulos” (Paso 3); en este ejemplo dejaremos el mismo alfa por defecto (0.05), excepto si amerita cambio (Paso 4) e indicamos la opción de salida de los resultados seleccionando la opción “Rango de salida” (preferencia personal del autor) (Paso 5) y definimos la celda donde se ubicará la tabla de resultados, es este caso en la celda E1 (Paso 6).

	A	B	C	D
1	Ríos	Alta	Media	Baja
2	Río Grande	5.6	4.3	3.2
3	Río Escondido	4.5	4.2	3.6
4	Río El Salto	4.2	3.8	3.2
5	Río Alegre	5.3	4.2	4.1
6	Río San Luis	2.1	3.4	2.9

Paso 1: Seleccionar 'Análisis de varianza de dos factores con una sola muestra por grupo' en el menú 'Análisis de datos'.

Paso 2: Seleccionar el rango de entrada '\$A\$1:\$D\$6'.

Paso 3: Marcar la opción 'Rótulos'.

Paso 4: Mantener el nivel de significancia 'Alfa' en 0.05.

Paso 5: Seleccionar la opción 'Rango de salida'.

Paso 6: Definir el rango de salida como '\$E\$1'.

Figura 48. Pasos para realizar un análisis de varianza de dos factores sin replicación. Notemos la estructura de tabla de datos donde se definen los niveles de cada factor.

Las tablas de resultados (Figura 49) incluyen la tabla de estadísticas descriptivas por factor (resumen) y la tabla del ANDEVA (análisis de varianza), para el cual fijaremos la atención en el valor de “p”. La tabla muestra dos valores de “p”, uno para “Filas” (Factor 2= Ríos: Río Grande, Río Escondido, Río El Salto, Río Alegre, Río San Luis) y otro para “Columnas” (Factor 1= Partes de la microcuenca: Alta, Media y Baja).

Dado $\alpha = 0.05$, el valor de “p” es mayor que 0.05 para “Filas” y “Columnas” (0.064 y 0.126 respectivamente), por lo que fallamos en rechazar las hipótesis nulas para los dos factores y concluimos, que no hay diferencias en los valores de Oxígeno Disuelto comparado entre “Ríos” y entre “Partes de la microcuenca”.

Análisis de varianza de dos factores con una sola muestra por grupo										
RESUMEN	Cuenta	Suma	Promedio	Varianza						
Rio Grande	3	13.1	4.3666667	1.4433333						
Rio Escondido	3	12.3	4.1	0.21						
Rio El Salto	3	11.2	3.7333333	0.2533333						
Rio Alegre	3	13.6	4.5333333	0.4433333						
Rio San Luis	3	8.4	2.8	0.43						
Alta	5	21.7	4.34	1.893						
Media	5	19.9	3.98	0.142						
Baja	5	17	3.4	0.215						
ANÁLISIS DE VARIANZA										
Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F				
Filas	5.689333333	4	1.422333333	3.4369714	0.064609043	3.837853355				
Columnas	2.249333333	2	1.124666667	2.7176802	0.125707925	4.458970108				
Error	3.310666667	8	0.413833333							
Total	11.24933333	14								

Figura 49. Resultados del análisis de varianza de dos factores sin replicación. Arriba se muestran las estadísticas descriptivas, abajo la tabla del análisis de varianza.

Relaciones entre variables

Determinar relaciones entre variables, es una de las tareas más comunes en estadística. Dos variables están relacionadas sí y solamente si los valores de una incrementan o disminuyen en función de los valores de la otra, es decir los valores de la variable hipotética “X” incrementan o disminuyen a la vez que los valores de la variable hipotética “Y” también incrementan o disminuyen. Por ejemplo, suponiendo que la variable “altura” de los individuos de una especie de árbol, está relacionada con la variable “elevación a nivel del mar”, en la cual los árboles más grandes ocurren a menor elevación y los árboles más pequeños se registran a mayor elevación. Podríamos decir que existe una correlación negativa entre las dos variables, en la cual los valores de la variable “altura” disminuyen al aumentar los valores de la variable “elevación”.

Sin embargo, las relaciones no determinan causalidad, es decir el hecho de que dos variables estén relacionadas, no significa que una sea la causante del incremento o reducción de los valores de la otra. Para el ejemplo de altura de árboles – elevación, puede que la elevación no sea la causa directa que los árboles sean grandes en los valles y pequeños en las cumbres, sino que otras variables como la “velocidad del viento” o la “pedregosidad del suelo” (por ejemplo), las cuales también varían a lo largo del gradiente de elevación, estén influyendo en la altura de los árboles.

Aplicaciones de Estadística Básica

Coefficiente de correlación

Las correlaciones se determinan y evalúan mediante el coeficiente de correlación, en este sentido se pueden obtener valores estadísticos que miden la significancia, fuerza y dirección de la relación. De manera general, y como primer paso, la relación se explora de forma gráfica utilizando un gráfico de dispersión de puntos. Si la relación existe, se observará una nube de puntos agrupados en una forma más o menos elíptica dirigida hacia la parte superior derecha o superior izquierda del gráfico a como se muestra en la figura 50.

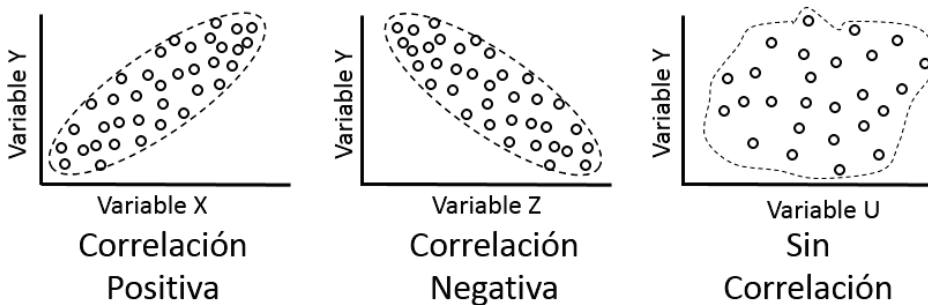


Figura 50. Ilustración de las relaciones y direcciones entre pares de variables. A la izquierda una relación positiva entre las variables “Y” y “X”; en el centro una relación negativa entre las variables “Y” y “Z”; a la derecha no se observa patrón de relación entre las variables “Y” y “U”.

Una correlación positiva indica que los valores de una variable incrementan, al incrementar los valores de la otra o disminuyen al disminuir los valores de la otra; una correlación negativa significa que los valores de una variable decrecen al incrementar los valores de la otra o viceversa; cuando no hay correlación, las variables se comportan de forma independiente.

Las pruebas de correlaciones asumen las siguientes hipótesis:

H_0 =No hay correlación significativa.

H_1 =Hay correlación significativa.

Para ejemplificar, utilizaremos datos hipotéticos de temperatura y elevación, y representaremos su relación mediante un gráfico de dispersión (personalizado), siguiendo la secuencia de opciones Insertar>Gráfico>Dispersión (Figura 51). Es notoria una relación negativa entre las dos variables.

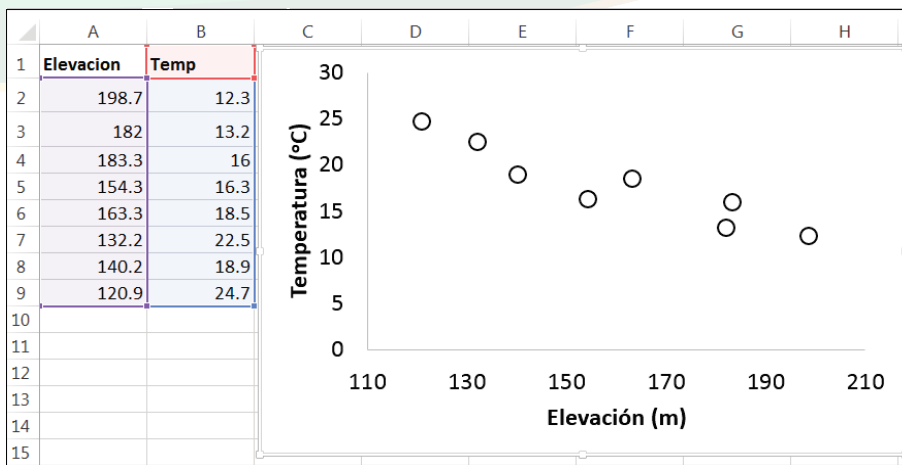


Figura 51. Dado los valores hipotéticos de dos variables ambientales “Elevación” y “Temperatura”, se observa una correlación negativa entre ellas.

Después de visualizar la relación necesitamos confirmar la significancia y fuerza de la relación, para lo cual utilizaremos la opción “Coeficiente de correlación”. La opción devolverá el Coeficiente de Correlación de Pearson el cual varía de 0 a 1, acercándose a 0 cuando la correlación es muy débil y acercándose a 1 cuando la correlación es muy fuerte. El mismo coeficiente confirma la dirección de la relación con un signo negativo (-) si la relación es negativa y sin signo, si la relación es positiva.

La opción la encontramos en la herramienta de “Análisis de datos” en “DATOS”, del cuadro de diálogo que se despliega seleccionamos la opción llamada “Coeficiente de correlación” (Paso 1) (Figura 52), luego seguimos los mismos procedimientos mostrados para los análisis anteriores, en el cual se selecciona el rango de los datos, haciendo clic en botón de selección (el de la flecha roja) (Paso 2), el cuadro de diálogo se minimizará y dará lugar a que podamos seleccionar los datos (las dos columnas con sus encabezados), seguidamente maximizamos el cuadro de diálogo haciendo clic de nuevo en el botón de selección (Paso 3) y completamos las otras opciones en el cuadro de diálogo, para este ejemplo pondremos el check donde dice “Rótulos en la primera fila” (que corresponde a los encabezados de las variables) (Paso 4), seleccionamos “Rango de salida” (Paso 5) y con el botón de selección activamos la celda, a partir de la cual se desplegará el cuadro de resultado (C1) (Paso 6), luego finalizamos con “Aceptar”.

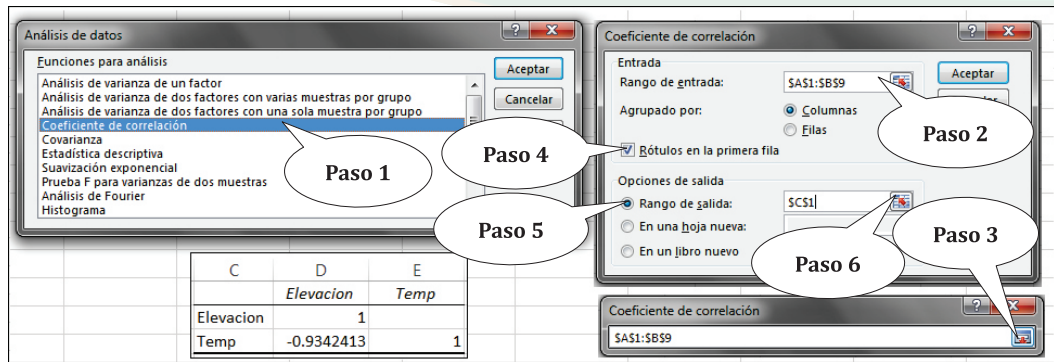


Figura 52. Ilustración de los pasos para realizar análisis de correlación en MS Excel. Se incluye la tabla de resultados.

La figura 52 muestra la tabla de resultado del cálculo del Coeficiente de Correlación de Pearson cuyo valor es -0.93 (redondeado a dos decimales). Este valor indica dos cosas, la dirección de la correlación, que en este caso es negativa y la fuerza de la relación, la cual como se acerca a 1 se concluye que la relación es muy fuerte.

En el capítulo relacionado con este mismo tema en la sección de “Estadísticas básicas en R”, se utilizará una función que no solamente mostrará el valor del coeficiente, sino un valor de “p” para determinar la significancia; adicionalmente, ofrece la opción de cambiar a una prueba no paramétrica como lo es el “Coeficiente de Correlación de Spearman”.

Regresión lineal simple

Cuando un par de variables están relacionadas, muchas veces es necesario obtener más información sobre esa relación, en especial es demandado determinar si la relación es lineal y, de serlo, obtener una fórmula para predecir los valores de una variable (dependiente) en función de la otra (independiente). Hay muchas formas de determinar cuál de dos variables es dependiente y cual es independiente. En principio, lo más intuitivo es determinar qué variable está influenciada por la otra, por ejemplo si al variar los valores de la variable X, se espera una respuesta de la variable Y, entonces X es la variable independiente y Y la variable dependiente; por el contrario, si al variar los valores de la variable X no se espera una respuesta de la variable Y, entonces es probable dos cosas: que la variable Y sea una variable independiente o que ambas variables sean independientes.

Regresando al ejemplo de la relación temperatura – elevación, hemos conocido que esta correlación es negativa, pero ¿Cuál es la variable dependiente y cuál la independiente?

Podríamos hacer el siguiente ejercicio de razonamiento: al aumentar la elevación, se esperaría (evidentemente), que la temperatura se reduzca, de igual forma al disminuir la elevación se esperaría que la temperatura incremente; lo contrario no es muy razonable: ¿Al aumentar o reducir la temperatura se esperaría aumento o disminución de la elevación? ¡La respuesta sería no! La temperatura alta o baja no afecta la elevación, la elevación alta o baja si afecta la temperatura, por tal razón definimos a la temperatura como la variable dependiente y a la elevación como la variable independiente. Para utilizar el análisis de regresión es necesario definir la variable dependiente e independiente, el no definirlas es probable que no tenga considerable impacto en los cálculos, pero sí en la interpretación final.

El análisis lo aplicamos con la opción “Regresión”, la cual se encuentra en Datos>Análisis de datos (Paso 1) (Figura 53); seguidamente seleccionamos los datos, primeramente se seleccionan los datos de la variable dependiente (Y, Temp) (Paso 2) y luego los de la variable independiente (X, Elevación) (Paso 3); ponemos check en la opción “Rótulos”, si en la selección se incluyeron los nombres de las variables (Paso 4); ponemos check en el “Nivel de confianza” y dejamos el mismo, o se cambia según convenga (Paso 5); luego seleccionamos la opción de salida, para el caso particular de este ejemplo se seleccionó la opción “Rango de salida” (Paso 6) y definimos la celda donde se pondrá la tabla de resultados dentro de la hoja de cálculo deseada (C1) (Paso 7). Seleccionamos los residuales, poniendo check en las opciones “Residuos” (Paso 8), “Gráfico de residuales” (Paso 9) y “Curva de regresión ajustada” (Paso 10). El resto de las opciones se dejarán a criterio del lector, según le convenga.

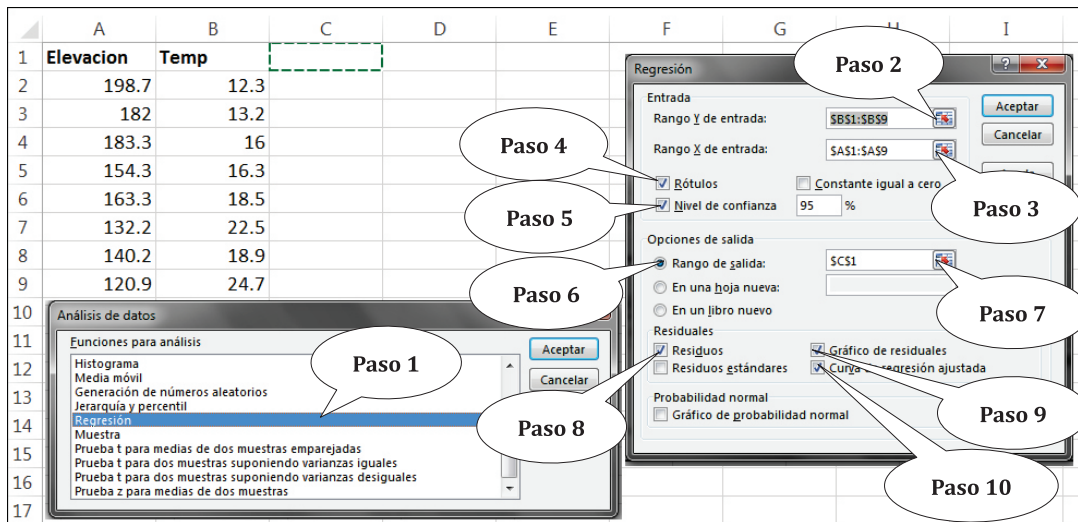


Figura 53. Ilustración de los pasos a seguir, para aplicar un análisis de regresión a los datos presentados en la figura 51.

Aplicaciones de Estadística Básica

Los primeros tres aspectos que debemos examinar en la tabla de resultados (Figura 54) de la regresión, son el coeficiente de determinación (R^2), el valor de “p” para F, y el valor y significancia de los coeficientes. El coeficiente de determinación (R^2) nos indica cuánto cambio en la variable dependiente ha sido explicado por la independiente (ajuste del modelo de regresión), o sea qué tanto de la información (que se usó en el cálculo de regresión) está siendo incluida en el modelo. Cuando R^2 se acerca a 1 significa que la mayoría de la información se incluyó en el modelo y que el modelo es certero y representativo, cuando R^2 se acerca a 0 significa que muy poca información se incluyó en el modelo y que dicho modelo no estaría representando la verdadera relación entre las dos variables. En el caso del ejemplo anterior $R^2 = 0.87$ (Examinar 1), lo que indica que el 87% de la variación en “Y” (Temperatura) es explicada por el modelo.

Resumen						Análisis de los residuales					
Estadísticas de la regresión						Observación		Pronóstico Temp		Residuos	
Coefficiente de correlación múltiple	0.9342413	Examinar 1				1		12.05946463		0.240535367	
Coefficiente de determinación R^2	0.8728068					2		14.49650157		-1.296501573	
R^2 ajustado	0.851608					3		14.30679211		1.69320789	
Error típico	1.6491858					4		18.53877243		-2.238772426	
Observaciones	8					5		17.22539922		1.274600776	
						6		21.76383329		0.736166713	
						7		20.59639044		-1.696390441	
						8		23.41284631		1.287153694	
ANÁLISIS DE VARIANZA											
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F						
Regresión	1	111.9811177	111.9811177	41.17234846	0.000676286						
Residuos	6	16.31888225	2.719813708								
Total	7	128.2999999									
	Coefficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%			
Intercepción	41.05582631	3.670942461	11.18400159	3.04982E-05	32.0733537	50.03829892	32.0733537	50.03829892			
Elevacion	-0.145930356	0.022742742	-6.416568278	0.000676286	-0.20157984	-0.090280871	-0.20157984	-0.090280871			

Figura 54. Resultados del análisis de regresión simple. Se muestra el coeficiente de determinación (Examinar 1), el valor de “p” para el modelo (Examinar 2), el valor de “p” para los coeficientes (Examinar 3) y la tabla de residuos (Examinar 4).

Si el valor de “p” para F (Examinar 2) es menor de 0.05, significa que al menos uno de los coeficientes es significativamente diferente a cero, y expresa que el modelo como un todo es significativo, lo contrario ($p \geq 0.05$) indica que ninguno de los coeficientes es diferente a cero y el modelo de regresión no es útil. En el caso del ejemplo, el valor de p (0.00068) es muy pequeño comparado con 0.05, por lo que concluimos que el modelo es significativo.

Los coeficientes son el intercepto y la pendiente (Elevacion) de la línea de regresión del modelo y el valor de p (probabilidad) para cada coeficiente (Examinar 3) determina las variables significativas a ser mantenidas en el modelo, si el coeficiente (de la variable independiente) no es significativo, se debería de tener cautela en el uso del mismo. En el

caso del ejemplo, el valor de p (0.00068) de la “Elevación” es muy pequeño, comparado con 0.05, con lo que concluimos que el coeficiente es significativo.

Los residuos se obtienen de la diferencia entre los valores observados y los valores predichos por el modelo de regresión, esta resta puede generar tanto valores positivos como negativos (Examinar 4). Dada la lista de observaciones y sus residuos, debemos identificar los valores muy altos o muy bajos en relación a cero, estos valores identificados están por encima o por debajo de los valores predichos y pueden ser potenciales valores atípicos (Figura 55).

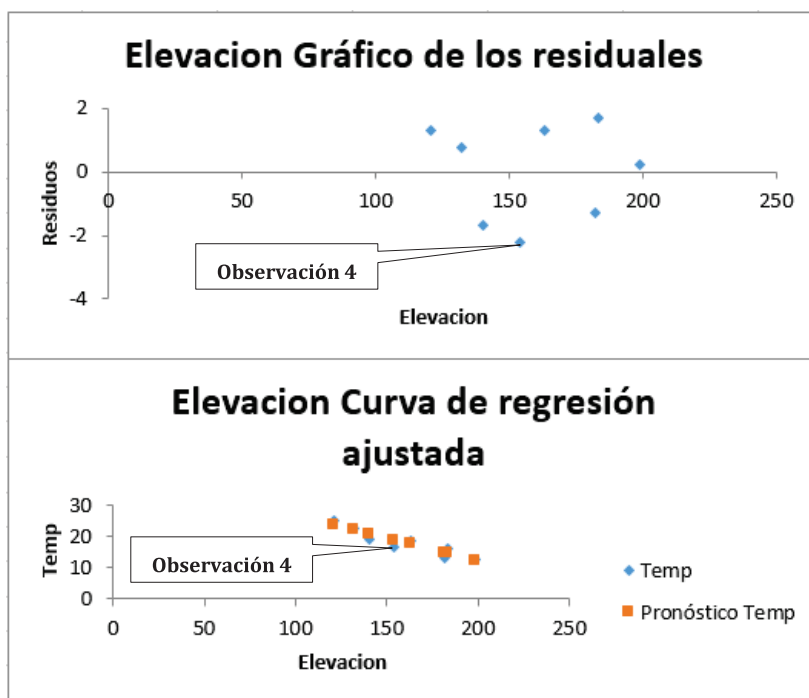


Figura 55. Arriba se presenta el gráfico de residuales y abajo el gráfico de regresión ajustada.

En la figura 55 notamos que entre los residuos con mayor valor se encuentra la “Observación 4” (residuo= -2.2388) (Figura 54). En el gráfico de arriba observamos los puntos dispersos uniformemente, con las distancias de los puntos a la línea de cero (0) casi similares arriba y debajo. Incluso el punto más distante a la línea 0 (Observación 4: residuo= -2.2388) no se observa alejado del resto de los puntos, ubicados en la parte negativa del eje de residuos, por lo que no se podría considerarlo un valor atípico.

Aplicaciones de Estadística Básica

En el gráfico de abajo, en la figura 55, se han superpuesto los valores observados de la variable dependiente (Temperatura) con los valores predichos, de tal forma que podemos observar una superposición casi perfecta, a excepción de algunos puntos que están un poco alejados de la línea que conforma los valores predichos, en particular un valor que corresponde a la “Observación 4”.

En la figura 56 se presentan seis ejemplos, que ilustran la distribución de los puntos en un gráfico de residuales, de los cuales solamente el primero (Figura 56 A) es el que muestra una distribución normal de los residuos; o sea, que los puntos se distribuyen uniformemente arriba y debajo de la línea punteada con origen en cero, y que, además, los puntos se observan a distancias regulares entre ellos. Los restantes cinco ejemplos (Figura 56 B – F) muestran situaciones atípicas, que nos pueden advertir sobre un problema en el modelo de regresión lineal o que los datos siguen otra forma de relación.

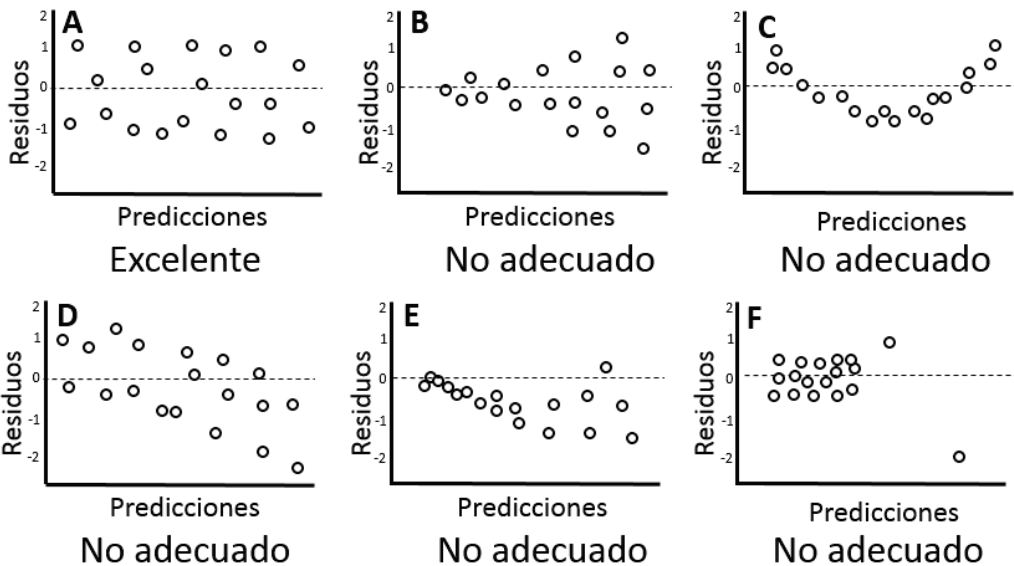


Figura 56. Ilustración de seis situaciones donde se visualizan los residuos en el gráfico de residuales. El panel A muestra la condición de excelente; de B a F se muestran varias situaciones no adecuadas para una regresión lineal.

Para realizar la predicción de la variable dependiente (Temperatura) sobre la base de los valores de la variable independiente (Elevación), se utiliza la fórmula:

$$Y = a + b (X)$$

Donde,

Y= Variable dependiente (a predecir)

a= Intercepción

b= Pendiente

X= Variable independiente (predictora)

Si, $a = 41.06$ y $b = -0.15$, la fórmula de predicción quedaría expresada como:

$$Y = 41.06 + (-0.15) (X)$$

Para usar la fórmula, supondremos, que, por ejemplo, se necesitaría predecir la temperatura (Y) cuanto la elevación (X) es 150 m, entonces se sustituye en la fórmula:

$$Y = 41.06 - 0.15 (150) = 18.56^{\circ}\text{C}$$

O sea, que a los 150 metros de elevación se esperaría una temperatura aproximada de 18.56°C .

Regresión lineal múltiple

El principio de la regresión lineal múltiple es similar al de la regresión lineal simple, solamente que en la múltiple se incluyen más de dos variables, de las cuales una es la variable dependiente (o predicha) y el resto son variables independientes (o predictores). Para ejemplificar la aplicación de este análisis, utilizaremos una matriz de datos formada por cuatro variables, una de las cuales es la variable dependiente, en este caso es la cobertura (%) de *Anomodon attenuatus* (una especie de musgo epífito) medida a 20 cm del suelo en la base de los árboles; y las variables predictoras son la temperatura del aire ($^{\circ}\text{C}$), la humedad relativa del aire (HR) y el diámetro del árbol hospedero (cm). Con el análisis se pretende predecir la cobertura del musgo con la combinación de valores de las otras variables.

En MS Excel, utilizaremos la misma opción para realizar una regresión lineal simple, solamente que incluiremos varias variables independientes, para ello seguiremos la secuencia de opciones Datos>Análisis de datos y se elige “Regresión” (Paso 1) (Figura 57); seguidamente seleccionamos los datos de la variable dependiente (Y, Cober) (Paso 2)

Aplicaciones de Estadística Básica

y luego los de las variables independientes (Temp, HR y DAP) (Paso 3), ponemos check en la opción “Rótulos” si en la selección incluimos los nombres de las variables (Paso 4), ponemos check en el “Nivel de confianza” y dejamos el mismo, o lo cambiamos según convenga (Paso 5), luego seleccionamos la opción de salida; para el caso particular de este ejemplo seleccionamos la opción “Rango de salida” (Paso 6) y definimos la celda donde el programa colocará la tabla de resultados dentro de la hoja de cálculo (E1) (Paso 7). Adicionalmente seleccionamos los residuales poniendo check en las opciones “Residuos” (Paso 8) y “Gráfico de residuales” (Paso 9). Hay más opciones que podemos seleccionar, pero para este ejemplo explorar esas opciones se dejará a criterio del lector.

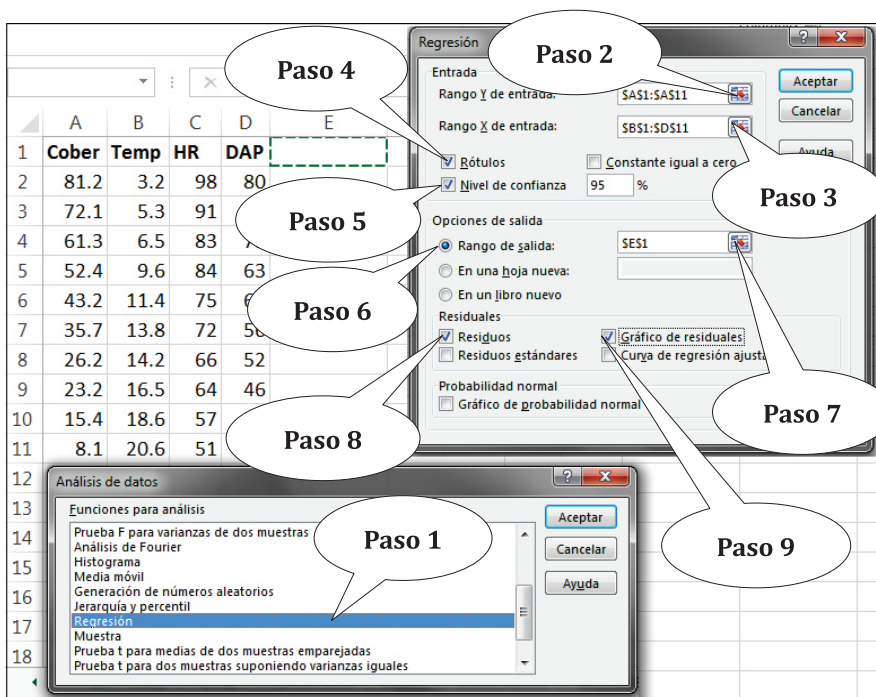


Figura 57. Pasos para realizar un análisis de regresión múltiple. Cober= cobertura del musgo en porcentaje; Temp= temperatura del aire en °C; HR= humedad relativa del aire en porcentaje; DAP= diámetro en cm de los árboles hospederos, medidos a la altura del pecho.

Los primeros tres aspectos que debemos examinar en la tabla de resultados (Figura 58) de la regresión, son el coeficiente de determinación (R^2), el valor de “p” para F, y el valor y significancia de los coeficientes. El coeficiente de determinación (R^2) (Examinar 1) indica el ajuste del modelo de regresión, en el caso del ejemplo $R^2 = 0.99$, lo que indica que el 99% de la variación en “Y” (Cobertura) es explicada por el modelo. El valor

de “p” (Examinar 2) es muy pequeño comparado con 0.05, con lo que concluimos que el modelo es significativo. El valor de “p” de las variables “Temp” (temperatura) y “HR” (humedad relativa del aire) son menores que 0.05, pero el valor de “p” de la variable “DAP” (diámetro de los árboles a la altura del pecho) es mayor que 0.05, por lo que solo mantenemos las variables “Temp” y HR en el modelo (Examinar 3). Y a criterio del lector, se podría excluir la variable “DAP”, aunque para ello sería conveniente aplicar ciertos procedimientos y así obtener evidencias objetivas sobre la inclusión o exclusión de la variable, estos procedimientos solamente son abordados en el programa R. Los valores atípicos podrían ser la observación 7 (-3.797...), la observación 4 (-2.839...) y la observación 6 (2.145...) por poseer los residuos más alejados de cero (Examinar 4).

				Análisis de los residuales			Examinar 4	
				Observación	Pronóstico	Cover	Residuos	
Resumen				1	80.91512457	0.28488		
				2	70.27669556	1.8233		
				3	60.87490869	0.42509		
Estadísticas de la regresión				4	55.23967316	-2.8397		
Coeficiente de correlación múltiple	0.9969864	Examinar 1		5	43.09991179	0.10009		
Coeficiente de determinación R^2	0.993982		6	33.55411018	2.14589			
R^2 ajustado	0.990973		7	29.99772438	-3.7977			
Error típico	2.3324731		8	23.23737916	-0.0374			
Observaciones	10		9	14.73247247	0.66753			
				10	6.872000052	1.228		
ANÁLISIS DE VARIANZA								
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F			
Regresión	3	5391.493415	1797.164472	330.3349574	4.75694E-07			
Residuos	6	32.64258471	5.440430785		Examinar 2			
Total	9	5424.136	Examinar 3					
	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
Intercepción	63.16353696	41.24891931	1.531277377	0.17658138	-37.76893254	164.0960065	-37.76893254	164.0960065
Temp	-3.857116594	1.171884965	-3.29137817	0.01658526	-6.724615803	-0.989617385	-6.724615803	-0.989617385
HR	0.862408687	0.321837758	2.679637998	0.036553789	0.074900061	1.649917312	0.074900061	1.649917312
DAP	-0.682427154	0.413796785	-1.64918428	0.150204426	-1.694951412	0.330097104	-1.694951412	0.330097104

Figura 58. Resultados del análisis de regresión múltiple. Se muestra el coeficiente de determinación (Examinar 1), el valor de “p” para el modelo (Examinar 2), el valor de “p” para los coeficientes (Examinar 3) y los residuales (Examinar 4).

En la figura 59 se muestran los gráficos de residuos para cada una de las variables independientes, si dichos gráficos son comparados con las diferentes situaciones mostradas en la figura 56, se concluye que no se observa una distribución fuera de lo normal de los residuos en los gráficos de residuales; a pesar de los potenciales valores atípicos (observaciones 7, 4 y 6).

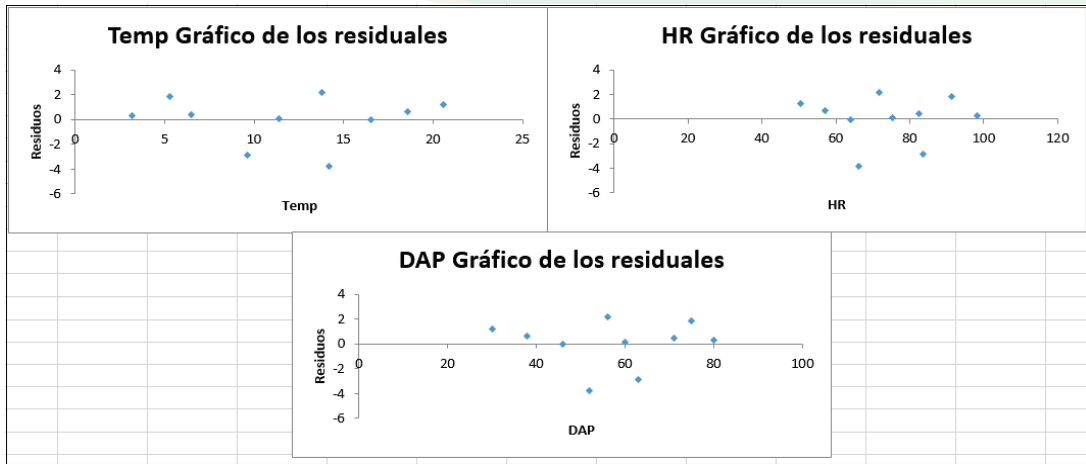


Figura 59. Gráficos de residuos resultantes del análisis de regresión múltiple, uno para cada variable independiente. Temp= temperatura del aire en °C; HR= humedad relativa del aire en porcentaje; DAP= diámetro en cm de los árboles hospederos medidos a la altura del pecho.

De tal forma que la predicción de la cobertura de la especie de musgo epífito se haría con la ecuación:

$$\text{Cobertura \%} = 63.16 - 3.86 (\text{Temperatura } ^\circ\text{C}) + 0.86 (\text{Humedad Relativa \%}) - 0.68 (\text{DAP cm})$$

En un lugar donde registremos una temperatura de 18 °C, una humedad relativa del aire de 87% y en un árbol de 80 cm de diámetro, se esperaría encontrar una cobertura de *Anomodon attenuatus* de:

$$\text{Cobertura} = 63.16 - 3.86(18) + 0.86(87) - 0.68(80)$$

$$\text{Cobertura} = 63.16 - 69.48 + 74.82 - 54.4$$

$$\text{Cobertura} = 14.1\%$$

Sobre las opciones no paramétricas

Si la transformación de los datos no es una opción (sea porque los datos continúan sin ajustarse a una distribución normal, incluso después de la transformación o porque el investigador simplemente no desee hacer transformaciones), entonces tendremos que recurrir a pruebas no paramétricas. Estas no requieren que los datos se ajusten a una distribución normal. Las pruebas transforman los datos en rangos y aplican diferentes algoritmos para lograr las correspondientes comparaciones.

Ninguna de estas pruebas se abordará para MS Excel, pues no están disponible de forma automática y sus abordajes manuales en las hojas de cálculo son tediosos. Estas pruebas se pueden aplicar instalando una extensión estadística a MS Excel o utilizando el programa R. En el cuadro 4 se presenta una lista de las pruebas paramétricas más populares y sus equivalencias no paramétricas.

Cuadro 4. Lista de las pruebas paramétricas más populares con sus pruebas no paramétricas equivalentes.

PRUEBA PARAMÉTRICA	ALTERNATIVA NO PARAMÉTRICA
Prueba T para una muestra	Prueba de Wilcoxon
Prueba T para dos muestras independientes	Prueba de Wilcoxon o Mann-Whitney
Prueba T para dos muestras pareadas	Prueba de Wilcoxon pareada
Análisis de Varianza de una o dos vías para muestras independientes	Kruskal-Wallis
Análisis de Varianza de una o dos vías para muestras repetidas	Prueba de Friedman
Coeficiente de correlación de Pearson	Coeficiente de correlación de Spearman
Regresión	Kendall's Tau (No abordado)

Tomado de: Kiernan & Bevilacqua (2011)

Opciones gráficas

Gráficos básicos

Microsoft Excel nos ofrece opciones gráficas muy versátiles y muy amigables para personalizar. Los gráficos básicos se ejemplificarán con el uso de tres matrices de datos hipotéticos (Figura 60), cada una de ellas servirá para generar un gráfico de barra, un gráfico de línea, un gráfico de pastel y un gráfico de dispersión.

La “Tabla de datos 1” tiene dos variables, una categórica nominal y una numérica continua, las cuales corresponden a datos de pH del suelo en cinco puntos, en un área determinada. Esta la representaremos por un gráfico de barras y un gráfico de línea. La “Tabla de datos 2” tiene dos variables, una categórica nominal y una numérica discreta, que contiene información sobre la cantidad de especies por clase taxonómica, para el grupo de Fauna Silvestre. La tabla va a ser representada por un gráfico de pastel. La “Tabla de datos 3” tiene dos variables numéricas continuas, que corresponden a la relación entre dos variables ambientales, temperatura (°C) y humedad relativa del aire (%) y va a ser representada por un gráfico de dispersión.

	A	B	C	D	E	F	G	H
1	TABLA DE DATOS 1			TABLA DE DATOS 2			TABLA DE DATOS 3	
2	Nº Punto	pH		Clases	#Especies		Temp	Humedad
3	Punto1	4.7		Aves	106		15.3	91.1
4	Punto2	5.3		Mamíferos	33		18.4	63.8
5	Punto3	5.9		Reptiles	27		22.5	32.7
6	Punto4	4.9		Anfibios	12		28	25.8
7	Punto5	4.6					34.7	20

Figura 60. Tres tablas de datos a ser utilizadas para demostrar la creación de gráficos básicos.

Gráfico de barras

Con la “Tabla de datos 1” elaboraremos un gráfico de barras, para ello tenemos que seleccionar el rango de datos de la tabla (Paso 1) y buscar las opciones gráficas en la pestaña “Insertar” (Paso 2), en donde seleccionaremos las opciones “Insertar gráfico de columna” (Paso 3) y “Columna agrupada” (Paso 4) (Figura 61).

Notemos en la misma figura 61 que se genera un gráfico automáticamente, el cual puede ser personalizado por el usuario. Por ejemplo, haciendo doble clic sobre el título iniciamos el modo de edición del mismo y podemos cambiar el nombre que se asigna por defecto (pH), o simplemente eliminamos el nombre y dejar un gráfico sin título. También podemos personalizar el color del fondo o borde de las barras, agregar tramas, cambiar grosor y distanciamiento con la opción “Dar formato a serie de datos”, que se despliega al hacer clic derecho sobre las barras (Paso 5), con ello se abre un panel de opciones donde se procede con la personalización (Paso 6). Las opciones de cambio del diseño del gráfico son múltiples en MS Excel, las podemos explorar en la pestaña “Diseño”, cuando el gráfico está activo (hacer clic sobre el gráfico para activarlo).

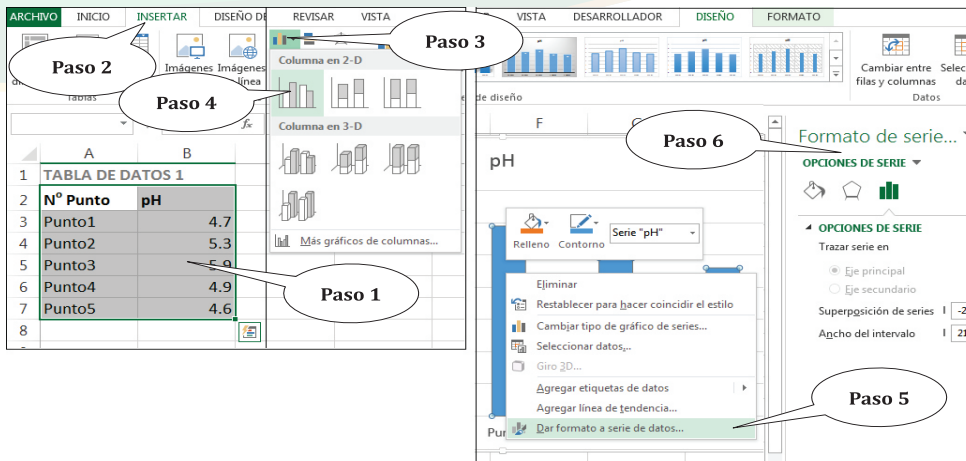


Figura 61. Ilustración de los pasos para crear y dar formato a un gráfico de barras.

Es imperativo que un gráfico tenga títulos de los ejes X y Y, para insertarlos, seleccionamos la opción “Agregar elemento de gráfico” (Paso 7) que se encuentra en la pestaña “Diseño” (Figura 62), luego elegimos “Títulos de ejes” y seleccionamos el eje que vamos a insertar. En principio, insertaremos el eje X (Horizontal primario) (Paso 8), notemos que en el gráfico aparece un recuadro de texto editable donde se escribe el nombre del eje (Puntos de muestreo) (Paso 9), luego insertamos el título del eje Y (Vertical primario) (Paso 10) y escribimos su nombre (pH). Logrando un gráfico de barra con sus correspondientes títulos de ejes.

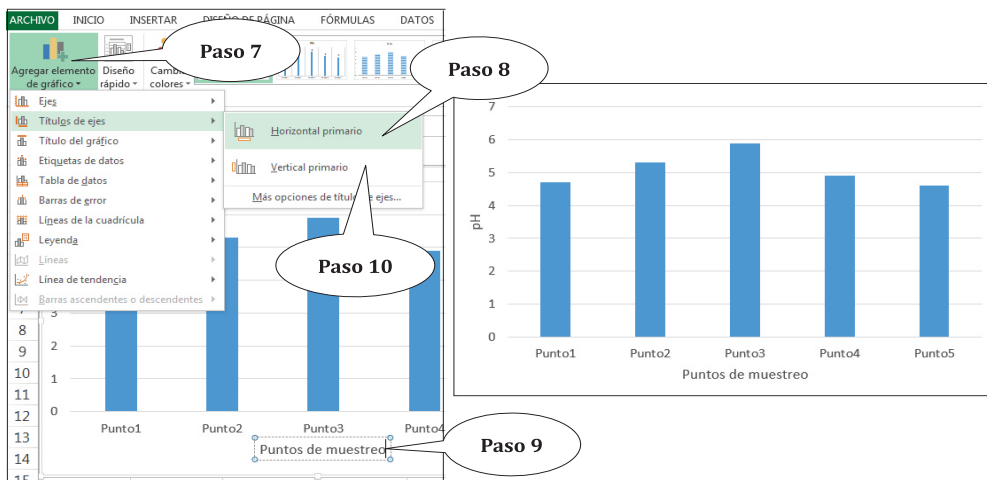


Figura 62. Ilustración de los pasos para insertar los títulos de los ejes X y Y. A la derecha se presenta el gráfico de barra sin título principal y con los títulos de los ejes.

Gráficos de área, líneas, pastel y dispersión

En la pestaña “Insertar” se presentan opciones para crear otros gráficos, entre ellos gráficos de área, líneas, pastel y dispersión, simplemente necesitamos seleccionar el tipo de gráfico requerido en el bloque de opciones llamado “Gráficos”. Si ponemos el puntero del mouse sobre cada uno de los íconos gráficos, el programa proveerá del nombre del gráfico y una descripción del mismo (Figura 63). Para crear un gráfico de área, nos dirigimos a la opción “Insertar gráfico de área”, para el gráfico de línea seleccionamos la opción “Insertar gráfico de línea”; para el gráfico de pastel buscar la opción “Insertar gráfico circular o de anillo”; para el de dispersión explorar la opción “Insertar gráfico de dispersión (X, Y) o de burbuja”.

La figura 63 muestra los cuatro gráficos creados con las tablas de datos en la figura 60. Todos los gráficos se pueden personalizar, de la misma forma que se personalizó anteriormente el gráfico de barras. Para aprovechar todas las opciones de personalización, se le invita al lector a que explore y se familiarice de forma personal con dichas opciones.

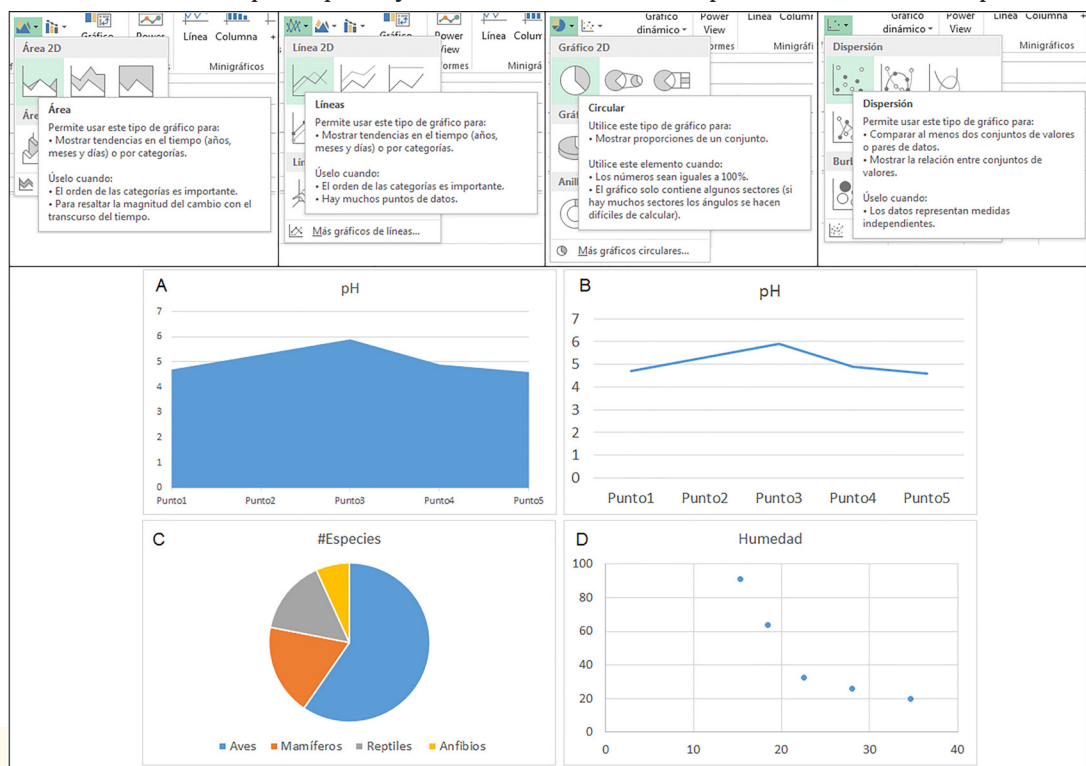


Figura 63. Opciones para crear gráficos de A. Área; B. Líneas; C. Pastel y D. Dispersión. El formato (colores y tamaños) y contenido (títulos principales y nombres de los ejes) que presentan los gráficos es el asignado por defecto por el programa.

Las barras de error

Incluir las barras de error en un gráfico de barras o de línea es algo muy común, esto lo podemos hacer fácilmente en MS Excel, sin embargo, requerirá ciertos pasos. En primer lugar, deberíamos tener una lista de medias asociada con una lista de descriptores de dispersión (desviación estándares o error estándar).

Ejemplificaremos con los datos de “pH” que se muestran en la figura 60, asumiendo que cada valor del pH en el ejemplo es una media que proviene de un conjunto de datos y para lo cual agregamos la lista de errores estándares (EE) a como se muestra en la figura 64. Después de haber agregado los datos del error estándar (Paso 1) y el gráfico (Paso 2), teniendo activo el gráfico nos dirigimos a “Diseño” (Paso 3) y seleccionamos la opción “Agregar elemento de gráfico” (Paso 4).

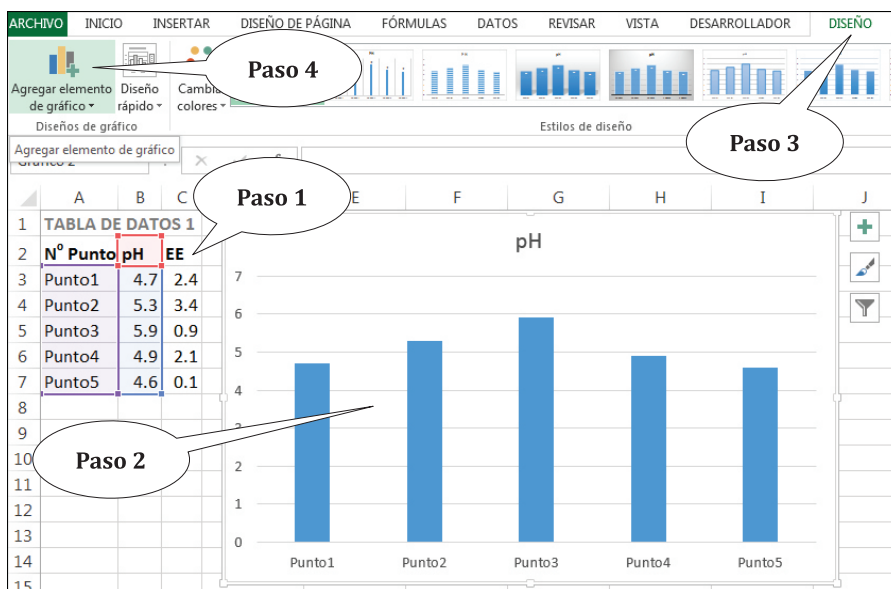


Figura 64. Esquematización de los pasos preliminares para insertar las barras de error en el gráfico de barras.

Al hacer clic en la opción “Agregar elemento de gráfico”, se despliegan todas las opciones que se pueden insertar en un gráfico para personalizarlo, en este seleccionaremos la opción “barras de error” y luego la opción “Más opciones de las barras de error...” (Paso 5) (Figura 65), ésta última opción nos permite el personalizar las barras de error.

Aplicaciones de Estadística Básica

Al seleccionar “Más opciones de las barras de error...” aparecerá una paleta de opciones en la parte derecha de la pantalla, la cual nos presenta todas las opciones para personalizar las barras de error. Para insertar barras de error tradicionales, seleccionaremos la opción que dice “Más” (Paso 6) y la opción llamada “Personalizado” (Paso 7), en esta última opción especificamos los valores de las barras de error, haciendo clic en “Especificar valor” (Paso 8).

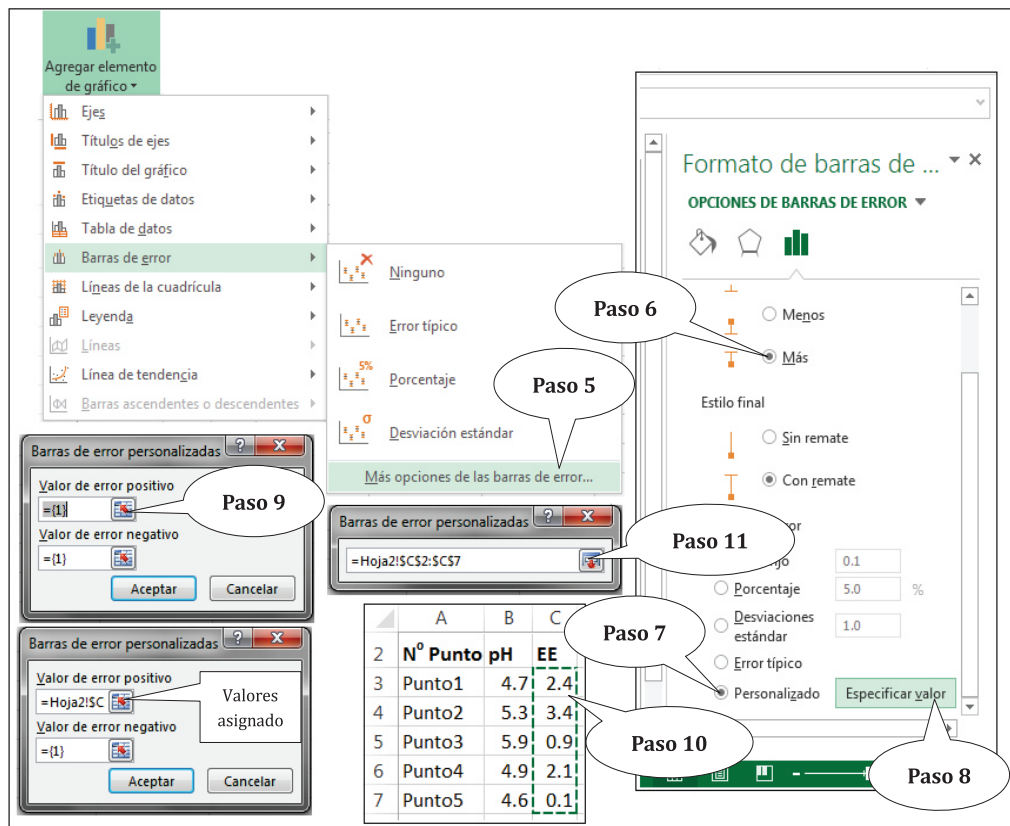


Figura 65. Ilustración de los pasos para insertar las barras de error a un gráfico de barras.

Aparecerá un cuadro de diálogo, donde especificaremos los valores de la barra de error, para ello hacemos clic en el botón con la flecha roja, donde dice “Valores de error positivo” (Paso 9) y el cuadro de diálogo se minimizará para darnos lugar a seleccionar los valores. Seguidamente seleccionamos los valores (solo los valores sin el encabezado de la columna) (Paso 10) y hacemos clic en la flecha roja del cuadro de diálogo de “Barras de error personalizadas” para maximizar dicho cuadro (Paso 11), se observa que el campo de “Valores de error positivo” tiene un rango de datos (=Hoja2!\$C....) que en conjunto

nos dice de dónde a dónde se ubican los valores (de la celda C3 a la C7 de la hoja 2) y finalmente, hacemos clic en “Aceptar”.

Notemos que en el campo de “Valores de error negativo” no hay valores porque, por el momento, solamente necesitamos la barra de error de arriba y no la de abajo. En caso que necesitaremos la barra de error de abajo, procederíamos a asignar los mismos valores de error estándar (EE) con el uso de la opción “Valores de error negativo” y los mismos procedimientos que usamos para el campo de “Valores de error positivo”.

El resultado es un gráfico de barras con barras de error hacia arriba (positivo) (Figura 66) el cual refleja el error estándar que se le asignó a los datos en la figura 64. De la misma forma se puede asignar el error estándar a otros tipos de gráficos, según lo amerite.

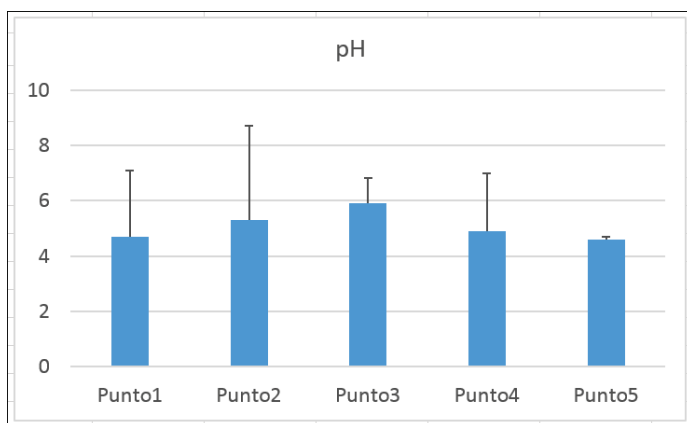


Figura 66. Gráfico de barra resultante para ejemplificar el uso de las barras de error.

La figura 66 muestra el gráfico de barras con barras de error a como típicamente se presentan en resultados de investigación, artículos y otras publicaciones científicas. Las barras de error nos pueden brindar ideas muy precisas de las diferencias entre las medias de cada conjunto de datos; sin embargo, es más fácil hacer interpretaciones cuando asignamos no solamente una de las barras de error, sino las dos barras de error, o sea la que se dirige hacia arriba (positivo) y la que se dirige hacia abajo (negativo).

Para lograr lo antes propuesto, seleccionamos las barras de error (solamente), hacemos clic derecho en una de ellas y escogemos la opción “Formato de barras de error...”, esta vez en la paleta de opciones se selecciona “Ambos”, donde dice “Dirección” (Paso 12); seleccionamos “Personalizado” y se especifican los valores (a como se realizó en los pasos 7 y 8 de la figura 65), seguidamente se especifican los valores para el error positivo (EE) y los mismos se especifican para el error negativo (Paso 13). El resultado se muestra en la figura 67.

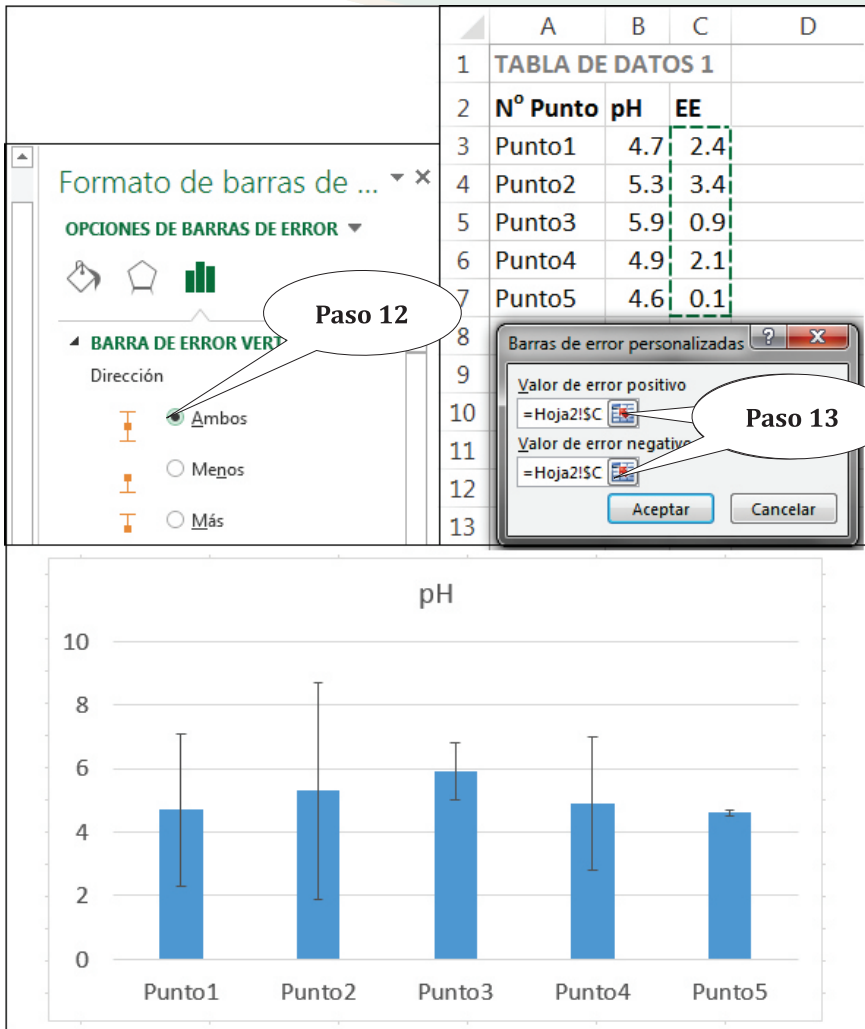


Figura 67. Ilustración de ediciones a la barra de error para insertar dos barras de error, una con orientación positiva y otra con orientación negativa.

La interpretación del gráfico resultante en la figura 67 es útil para explorar posibles diferencias significativas entre los conjuntos de datos. La regla general es que si se visualiza una línea imaginaria entre los extremos (positivos y negativos) de las barras de error de dos observaciones hacia la derecha y a la izquierda (de cada barra respectivamente), habrá una zona donde las áreas de ambas líneas imaginarias se superponen, si dentro de esa superposición quedan las medias de las dos observaciones entonces se sospecha que no haya diferencias significativas entre ambas medias (Figura 68 A).

En caso que el área resultante de la superposición de líneas imaginarias hacia la derecha e izquierda sea muy pequeña y que por lo tanto no incluyan a las medias, aún se sospecharía que las dos medias no tienen diferencias significativas (Figura 68 B). Sin embargo, si las líneas horizontales imaginarias que parten de los extremos de las barras de error de ambas observaciones, no se encuentran, se tendrían fuertes sospechas que las medias difieren significativamente (Figura 68 C).

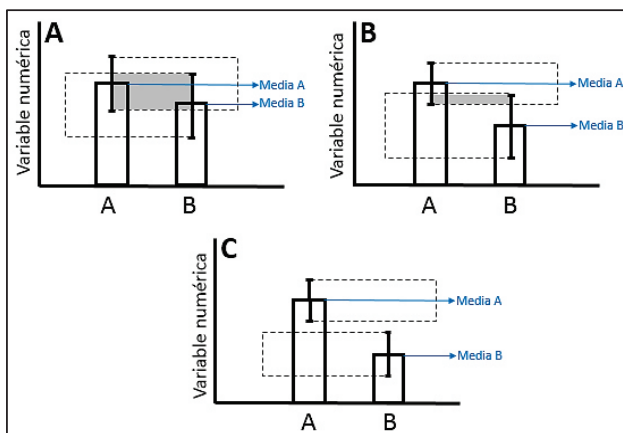


Figura 68. Ilustración de tres situaciones para interpretar las barras de error de dos direcciones. Las líneas imaginarias horizontales (líneas discontinuas) que parten de los extremos de la barra de error izquierda hacia la derecha y de los extremos de la barra de error derecha hacia la izquierda forman un área de superposición (área de color gris). A. Las medias están dentro del área superpuesta; B. El área superpuesta es muy pequeña y las medias no están incluidas dentro del áreas superpuesta; C. No existe área superpuesta.

Otros gráficos

Barras apiladas

Las barras apiladas son especialmente importantes para visualizar un conjunto de valores numéricos que se distribuyen en una variable categórica con dos niveles. Por ejemplo, en una encuesta cualquiera en la que se la hacía una pregunta a diferentes grupos de participantes, se determinaron las respuestas “SÍ” o “NO”; en cuyo caso podemos hacer un gráfico con los grupos y las categorías “SÍ” y “NO” más los valores de las frecuencias resultantes de la encuesta, los datos se presentan en la figura 69.

Aplicaciones de Estadística Básica

Para insertar el gráfico, seleccionamos todo el conjunto de datos (incluyendo los encabezados de columnas) y seleccionamos la opción de “barras apiladas” en “Insertar” (Figura 69). El gráfico quedará creado y cada color hará la diferencia entre los valores de respuesta “SÍ” y “NO” para completar un 100 % de frecuencias por barra. El paso siguiente correspondiente a la personalización del gráfico (cambio de colores, inserción de títulos de ejes, tipo/tamaño de letras, etc.) quedará en manos del lector.

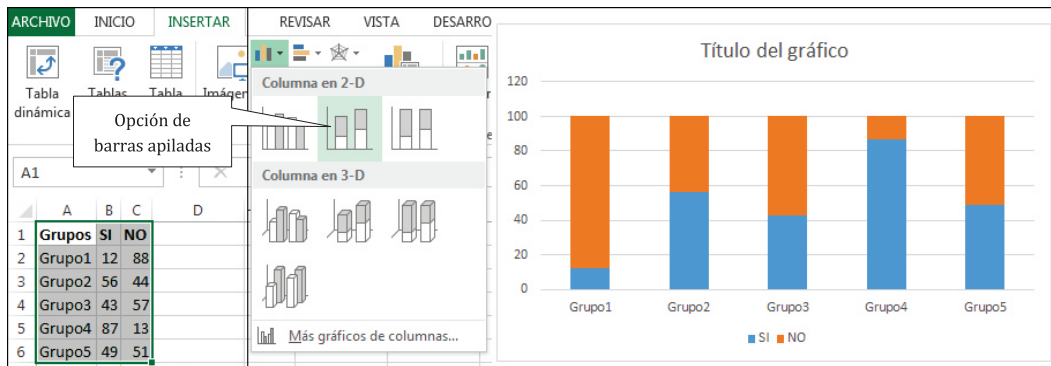


Figura 69. Ilustración del procedimiento para insertar un gráfico con barras aplicadas. En el ejemplo se encuestaron a cinco grupos de personas, cada grupo independiente del otro.

Eje X con dos o más variables categóricas

En general cuando se quieren representar datos numéricos, divididos por dos o más categorías, la representación de dichas categorías se logra con una adecuada estructuración de la tabla de datos. Por ejemplo, utilizando los mismos datos empleados para crear la figura 69, agregamos una nueva variable categórica “Sitio” con dos categorías, “Sitio1” y “Sitio2” a la izquierda de la variable “Grupos” y con ella insertamos el gráfico de barras, siguiendo los mismos procedimientos descritos en la figura 61. Notemos, si, que las columnas de las variables “Sitio” y “Grupo” no tienen encabezados (Figura 70) y que las categorías “Sitio1” y “Sitio2” no se repiten en todos los registros, solamente se colocan al inicio de cada conjunto de “Grupo”, perteneciente a cada sitio, ejemplo: en el Sitio1 se encuentra el Grupo1 y Grupo2; y en el Sitio2 se encuentran el Grupo3, Grupo4 y Grupo5.

En la figura 70 se muestra la nueva tabla de datos y el nuevo gráfico, notar que en el eje X se muestran las dos variables categóricas y que las categorías de la variable “Sitio” agrupa a las categorías de la variable “Grupo”.

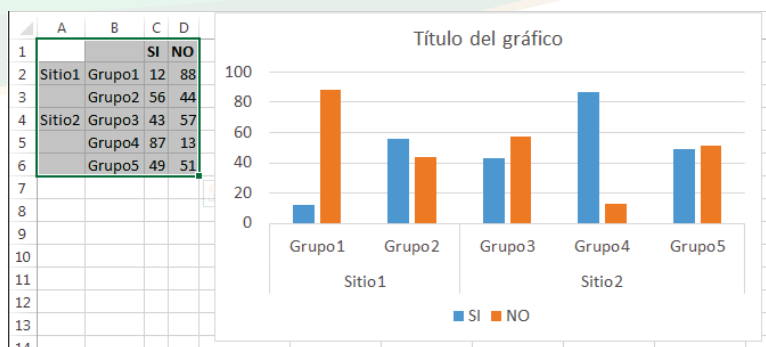


Figura 70. Ilustración de datos arreglados para generar un gráfico con dos variables categóricas (cada una con diferentes niveles de categorías) en el eje X.

Gráficos combinados

A como su nombre indica, los gráficos combinados resultan de la creación de un solo gráfico a partir de la combinación de dos gráficos. En general, se utilizan cuando deseamos representar la interacción de dos variables. Ejemplificaremos su uso con un pequeño conjunto de datos que relacionan la frecuencia de ocurrencia de incendios forestales con la temperatura media del ambiente. En este caso, tenemos en contexto dos variables, una variable numérica discreta que corresponde al número de incendios y una variable numérica continua correspondiente a la temperatura en °C. Asumimos que los datos se tomaron en cinco sitios, en cada uno se midieron la cantidad de incendios ocurridos y se registró la temperatura ambiental promedio antes de cada incendio.

Primeramente seleccionamos el conjunto de datos (Paso 1) (Figura 71), luego seleccionamos la opción “Crear gráfico combinado personalizado...” que se encuentra en la pestaña “Insertar” (Paso 2).

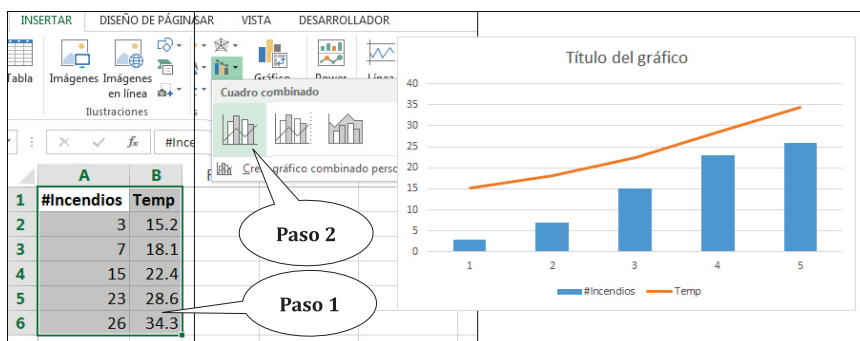


Figura 71. Representación del procedimiento para insertar un gráfico combinado. A la derecha se muestra el gráfico insertado.

Aplicaciones de Estadística Básica

Microsoft Excel asigna un tipo de gráfico para cada variable, en principio para la variable “#Incendios” el programa la presenta por defecto en un gráfico de barras y la variable “Temp” (Temperatura °C) la presenta en un gráfico de línea; sin embargo, estos pueden ser cambiados y personalizados. En la figura 72 se muestra el procedimiento para personalizarlo. Después de activar el gráfico (haciendo clic sobre él), seleccionamos la opción “Cambiar tipo de gráfico” (Paso 3) que se encuentra en la pestaña “Diseño”, la cual despliega un cuadro de diálogo con la opción llamada “Tipo de gráfico” que permite cambiar el tipo de gráfico a cada variable (Paso 4).

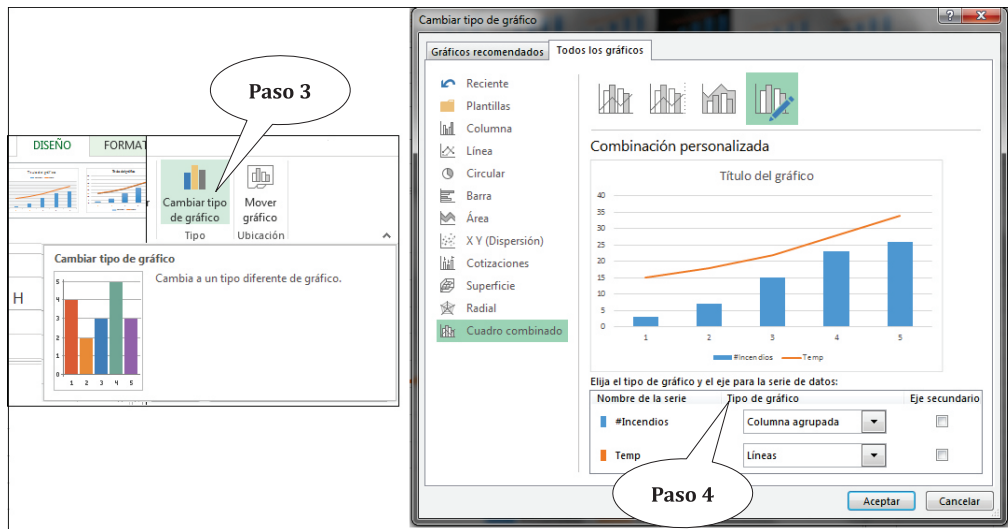


Figura 72. Representación del procedimiento para cambiar los tipos de gráficos para cada variable.

Gráfico de doble eje Y

Aunque no son muy comunes, en algunos momentos tendremos que crear un gráfico de doble eje Y. Este representa la relación de dos variables numéricas, una de las cuales se establece en el eje principal (eje Y de la izquierda) y otra en el eje secundario (eje Y de la derecha). Para ejemplificar la utilidad de este gráfico, emplearemos los datos del ejemplo anterior. Este tipo de gráfico lo creamos seleccionando la secuencia de opciones Insertar>Gráfico>Crear gráfico combinado personalizado..., en la cual elegimos la opción llamada “Columna agrupada – Línea en eje secundario” (Paso 1) (Figura 73). Adicionalmente añadimos los títulos de todos los ejes para diferenciar los dos ejes Y, esto lo logramos seleccionando Diseño>Añadir elemento de gráfico (Paso 2), añadimos el título horizontal primario (Sitios) y vertical primario (Número de incendios) así como fue mostrado en la figura 62 y luego añadimos el “Vertical secundario” (Temperatura °C) (Paso 3).

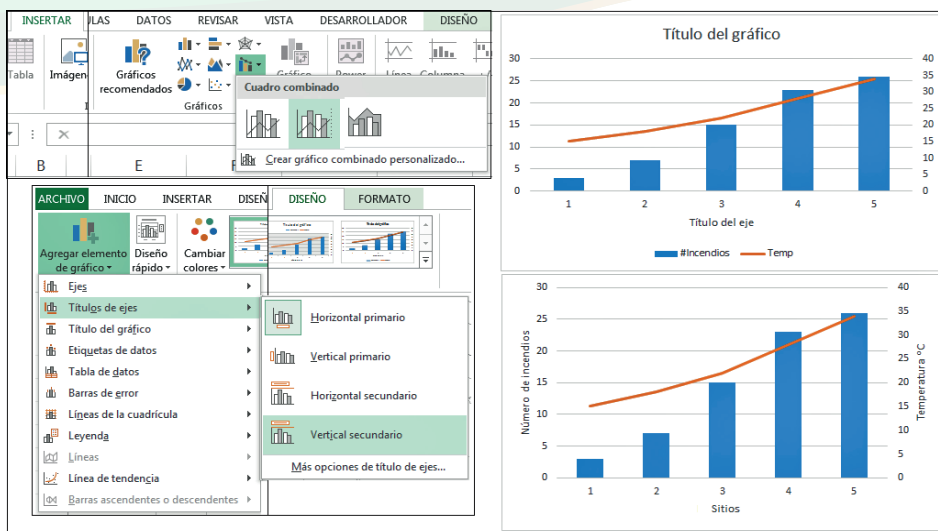


Figura 73. Pasos para crear un gráfico con doble eje Y. A la derecha, abajo se presenta el gráfico resultante.

Los gráficos creados automáticamente por MS Excel para cada variable, se pueden cambiar utilizando la opción “Cambiar tipo de gráfico”, descrita en la figura 72. De igual forma, en la misma figura se muestra cómo se puede reasignar el eje secundario a cualquiera de las dos variables, simplemente poniendo check en la variable que deseamos que aparezca en el eje secundario, en la opción llamada “Eje secundario”.

Gráficos miniatura

Los gráficos miniatura son pequeños gráficos que se insertan en una o algunas celdas de la hoja de cálculo. En general, no se suelen usar como gráficos formales para presentación de resultados a como se utilizan los anteriores, sino para representar gráficamente y rápidamente el comportamiento de los valores de un conjunto de datos.

Para ejemplificar su aplicación, utilizaremos unos datos de registro de crías por nidos, de una especie en particular de ratón de campo. En cinco sitios diferentes se registraron cinco nidos seleccionados de forma aleatoria y se contaron el número de crías. Se desea ver rápidamente las diferencias del número de crías entre los sitios y los nidos.

La información arreglada en una hoja de cálculo de MS Excel se muestra en la figura 74. Para insertar los gráficos miniaturas elegimos la opción Insertar>Minigráficos>Columna (Paso 1). Aparecerá un cuadro de diálogo donde especificaremos el “Rango de datos” y seleccionamos todos los datos de la tabla (B2:F6), evitando los encabezados

Aplicaciones de Estadística Básica

de columnas y nombres de filas (Paso 2), consecutivamente seleccionamos el rango de celdas donde aparecerán los gráficos miniatura (G2:G6) para este ejemplo (Paso 3) y seleccionamos “Aceptar”.

Paso 1

Paso 2

Paso 3

Minigráfico de columnas
Inserta un gráfico de columnas en una sola celda.

Crear grupo Minigráfico

Elija los datos para el grupo de minigráficos

Rango de datos: B2:F6

Elija la ubicación donde se colocará el grupo de minigráficos

Ubicación: \$G\$2:\$G\$6

Aceptar Cancelar

	A	B	C	D	E	F
1	Sitio	Nido1	Nido2	Nido3	Nido4	Nido5
2	Sitio1	3	2	1	3	0
3	Sitio2	2	3	2	3	5
4	Sitio3	1	2	1	0	3
5	Sitio4	6	3	4	2	7
6	Sitio5	0	0	1	2	1

Figura 74. Ilustración de la inserción de gráficos miniaturas. A la izquierda, abajo se muestra la tabla de datos con las gráficos miniaturas, en el rango de celdas G2:G6.

Estadísticas básicas en R

R Statistic es un programa estadístico muy versátil, flexible, de uso libre y con sólidas referencias a nivel global. El programa R fue creado por Ross Ihaka y Robert Gentleman en la Universidad de Auckland, Nueva Zelanda y actualmente desarrollado por el Equipo Principal de Desarrollo. El programa se llama simplemente “R”, por una parte aludiendo a las primeras letras de los primeros nombres de los autores (Ross y Robert) y por otra parte está relacionada con una historia del Laboratorio Bell, que creo el lenguaje S y del cual se deriva el lenguaje R, en cuyo caso “R” indica la letra antes de “S”. R está disponible para ser descargado e instalado desde la dirección electrónica: <http://cran.r-project.org/>

R puede realizar desde análisis estadísticos sencillos, hasta la creación de modelos muy complicados. Para ello, sus contribuidores han creado lo que en Microsoft Excel (MS Excel) llamamos “Complementos”, pero en R llamados “Paquetes”, que sirven para realizar análisis y visualizaciones muy específicas. Hay paquetes para diferentes áreas del conocimiento. Sin embargo, R trae su lista de opciones (funciones) de análisis “por defecto”, a lo que llamaremos “las funciones de R básico”, referido específicamente a estadística y graficación básica, por lo que en este escrito solo se hará referencia a los paquetes para ilustrar análisis muy particulares.

A diferencia de MS Excel, R es un programa estadístico con lenguaje de programación. En R no haremos referencia directa a las tradicionales filas, columnas y celdas como en MS Excel, sino se utilizará otro lenguaje. El lenguaje de programación y ambiente de R no es muy atractivo para los novicios y menos aún para los que prefieren programas donde simplemente las opciones se seleccionan con el uso de un “clic”.

R mostrará ciertos retos para los que se inicien en el uso de este programa; sin embargo, con la práctica se llegará a la conclusión de que es un programa más sencillo de lo que uno puede imaginar. Hay muchas cosas que lo hacen sencillo: en primer lugar el lenguaje de programación no es muy complicado y los códigos son predecible, en especial para las personas que tiene conocimientos básicos de la lengua inglesa; adicionalmente los códigos están totalmente disponibles en el programa, y si hay algo no disponible en el programa, existe toda una comunidad internacional que seguramente tiene códigos para todo tipo de análisis; gracias a la popularidad de R, y a esa comunidad internacional, es fácil encontrar ejemplos de aplicaciones de casi todos tipo de análisis y una amplia documentación en Internet.

Las aplicaciones de R para estadísticas y para graficar son increíblemente vastas y en este escrito es de interés solamente el abordar aspectos básicos, a fin de animar al lector a seguir explorando este mundo interminable de R Statistic.

Aplicaciones de Estadística Básica

Nociones sobre el ambiente de R

Primeros pasos en R

Después de instalado, al abrir R se desplegará la consola de R y presentará un mensaje de inicio de sesión (Cuadro 7), el cual se puede leer o se puede ignorar; lo que sí nos interesará en el símbolo mayor (>) y una barra vertical (|) intermitente, que es donde iniciaremos a escribir dentro de la consola del programa.

Cuadro 7. *Mensaje de inicio del programa R. Al final del mensaje se muestra el símbolo de mayor (>) que orienta al usuario donde se inicia la escritura en el programa.*

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>|
```

Iniciaremos el uso de R con acciones sencillas, primero utilizaremos el programa como una calculadora para hacer operaciones matemáticas básicas como sumar, restar, multiplicar y dividir. Entonces, a la par del símbolo ">" representado en el cuadro 7, escribiremos 2+2 y accionamos Enter; esto generará los números [1] 4, del cual el número [1] indicará el número de línea de la respuesta y el número 4 es la respuesta de la operación 2+2. En el programa se vería de la siguiente manera:

```
> 2+2
[1] 4
```

La consola de R trabaja por líneas, en una línea se escribe la operación y en otra genera el resultado (Figura 75) y después de haber producido el resultado, el indicador de una nueva línea de operación (“>”) aparece automáticamente, simbolizando que el programa está listo para recibir instrucciones para realizar otra operación.

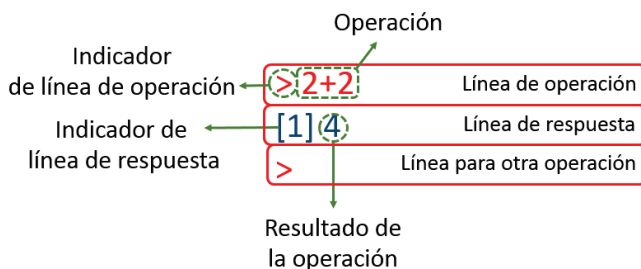


Figura 75. Representación de la línea de operación y línea de respuesta con sus componentes.

En R, el poner o no poner espacios entre los componentes de una operación no cambia el resultado, el programa reconoce los espacios y los caracteres por individuales de tal forma que escribir 2+2 sin espacios adelante y atrás del símbolo “+” generará idéntico resultado, que poniéndole los espacios en esas posiciones, o sea 2 + 2. En el programa se verificaría de esta forma:

```
> 2+2
[1] 4
> 2 + 2
[1] 4
```

Notemos que a cada suma 2+2 y 2 + 2 el programa las reconoce como operaciones separadas, por lo cual este produce una nueva línea de respuesta indicada por [1].

También podemos ejecutar otras operaciones matemáticas, entre ellas resta, divisiones y multiplicaciones. Por ejemplo, si escribimos 3-2 y presionamos Enter, obtenemos como respuesta 1; si escribimos 4/2 obtendremos como respuesta 2; si escribimos 2*2 obtenemos como respuesta 4. Observemos que para la operación de resta utilizamos el guion (-); para la operación de división utilizamos la pleca hacia la derecha (/) y para la operación de multiplicación utilizamos el asterisco (*). En R, estas operaciones se verían así:

Aplicaciones de Estadística Básica

```
> 3-2
[1] 1
> 4/2
[1] 2
> 2*2
[1] 4
```

En cada línea también podemos efectuar múltiples operaciones, por ejemplo podemos sumar los números 3, 2 y 4; luego a su resultado le restamos 6 y seguidamente lo multiplicamos por 3:

```
> 3+2+4
[1] 9
> 3+2+4-6
[1] 3
> (3+2+4-6) *3
[1] 9
```

La respuesta de $3+2+4$ es 9; la respuesta de $3+2+4-6$ es 3 y la respuesta de $(3+2+4-6)*3$ es 9. En esta última, el programa multiplica el número 3, con el resultado de la operación dentro del paréntesis. Esto indica que el uso de paréntesis en R para separar operaciones es tan crucial como para en MS Excel. Si la operación fuera ejecutada sin los paréntesis, la respuesta hubiera sido -9, ya que el programa estaría multiplicando primeramente el número 3 con el número -6 y a ese resultado se sumarían los restantes valores, esto sería un error de parte del usuario:

```
> (3+2+4-6) *3
[1] 9
> 3+2+4-6*3
[1] -9
```

Para la suma, en específico, R ofrece una función llamada “sum()”, de tal forma que solamente bastará escribir entre el paréntesis los números a sumar separados por coma y la suma se realizará:

```
> sum(3, 2, 4)
[1] 9
```

Para seguir haciendo pruebas de familiarización, podemos restar 6 al resultado anterior y posteriormente multiplicarle 3:

```
> sum(3, 2, 4) - 6
[1] 3
> (sum(3, 2, 4) - 6) * 3
[1] 9
```

En la primera línea de códigos utilizamos la función `sum()` para sumar los números 3, 2 y 4 en lugar de hacerlo con el símbolo “+”; seguidamente reescribimos toda la operación y le agregamos -6 para restarle el número 6 al resultado obtenido, generando el número 3 como resultado; finalmente hicimos la misma operación, pero le multiplicamos el número 3, colocando toda la operación de suma y resta entre paréntesis, de esta forma el programa puede multiplicar el número 3 con el resultado de la suma y resta.

Otras funciones matemáticas son `log()` que devuelve el valor de \ln (logaritmo natural o neperiano); `log10()` que devuelve el valor del logaritmo base 10; `logb()` presenta el valor de un logaritmo de base personalizada, el argumento sería `logb(X,b)` donde X es el número a determinar el logaritmo y b es el número que representa la base; `exp()` calcula el logaritmo con base “e”; `sqrt()` devuelve la raíz cuadrada; entre otros. Ejemplificaremos aplicando cada función al número 4, cuando sea necesario agregaremos un comentario al final de las funciones, en R los comentarios se insertan anteceditos por el símbolo numeral (#), estos no tienen ninguna influencia en las operaciones.

```
> log(4)
[1] 1.386294
> log10(4)
[1] 0.60206
> logb(4,3) # Devuelve el logaritmo de 4 con base 3.
[1] 1.26186
> exp(4)
[1] 54.59815
> sqrt(4)
[1] 2
```

Con lo anterior visto, deducimos que en R también podemos trabajar con fórmulas, para ejemplificar utilizaremos la fórmula del cálculo del área de un círculo de 25 m de radio, sabiendo que la ecuación es:

$$A = \pi r^2$$

Aplicaciones de Estadística Básica

Donde,

A = El área a calcular

π = 3.1416

r = Radio del círculo

Sustituyendo los valores en la fórmula, el área sería calculada como:

$$A = 3.1416 (25\text{m})^2 = 1963.5 \text{ m}^2$$

En R representaríamos la ecuación de la siguiente forma:

```
> 3.1416*(25^2)
[1] 1963.5
```

Notamos que es necesario indicarle al programa la separación de las operaciones, de tal forma que primero eleva el número 25 al cuadrado y luego que el resultado lo multiplique con el valor 3.1416, el resultado de la operación conjunta es el correcto y se muestra en la línea de respuesta [1].

El símbolo “^” le indica al programa que el número que está a la izquierda de símbolo (25) se va a elevar a la potencia y que el número que está a la derecha, es el valor de la potencia (2) al igual que en MS Excel.

Podemos guardar la respuesta de toda operación en un objeto virtual al que llamaremos “variable” (los objetos pueden ser números, letras, palabras, expresiones, gráficos, etc.), cuyo nombre es asignado por el usuario. Por ejemplo, asignaremos el resultado de la ecuación “3.1416*(25^2)” a una variable llamada “Area” (evitaremos el acento de la palabra momentáneamente), entonces cada vez que llamemos la variable “Area” (escribiendo su nombre en la consola de R y presionando Enter) R mostrará el valor resultante de la operación (1963.5). La asignación de la respuesta a la variable la realizamos con el operador de asignación representado por el signo menor (<) y un guion (-), o sea “<-” (también se puede usar el símbolo =), de tal forma que lo antes descrito quedaría representado en R de la siguiente forma:

```
> Area <-3.1416*(25^2)
```

A primera impresión no se muestra ningún resultado, porque el resultado solamente se mostrará al pedirle al programa que muestre el contenido de la variable “Area” y esto lo realizamos escribiendo en la consola el nombre de la variable y presionando Enter:

```
> Area <-3.1416*(25^2)
```

```
> Area # Para ver el contenido de la variable "Area" se presiona Enter.  
[1] 1963.5
```

Es importante no confundir el término de “variable”, como la designación donde guardamos objetos al término “variable” en estadística (atributos medibles), en cuyo caso los atributos medibles los podemos guardar como objetos en una variable en la consola de R. Por ejemplo, mediciones de la variable “Longitud de las Alas” de varios individuos de una especie de ave, pueden ser guardadas en la consola de R, en la variable llamada “Long_Alás”; notemos el doble uso de la palabra “variable”.

El formato vectorial

Una forma más avanzada de operar R es con el uso de formas vectoriales, definiéndose como vector en R como un conjunto de datos (caracteres numéricos o no numéricos) guardados en una variable. Por ejemplo, los números 3, 2 y 4 los podemos guardar en un vector con nombre “Var1”, la abreviación “Var” se refiere la palabra “Variable” (por ejemplo). Para guardar los números en el vector de nombre “Var1” haremos uso del operador de asignación y de la función de combinar “c()” a como se muestra a continuación:

```
> Var1 <-c(3,2,4)
```

Con ello los valores de 3, 2 y 4 han quedado guardados dentro de la variable “Var1”, para ver los valores escribimos el nombre de la variable (exactamente a como lo escribimos al inicio) y presionamos Enter.

```
> Var1  
[1] 3 2 4
```

En la figura 76 se explica el nombre de cada uno de los componentes:

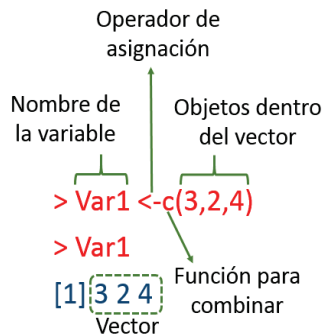


Figura 76. Representación de los componentes para asignar los valores a una variable a fin de estructurar un vector.

Aplicaciones de Estadística Básica

También podemos emplear los vectores en operaciones matemáticas, por ejemplo podemos sumar el número 2 a todos los valores de la variable “Var1” ejecutando la operación “Var1+2” sin necesidad de escribir los números que están dentro de la variable:

```
> Var1+2  
[1] 5 4 6
```

Para ejecutar la operación “Var1+2”, el programa le sumó el número dos a cada valor: 3+2, 2+2, 4+2 lo que resultó el 5, 4 y 6. Así podemos realizar otras operaciones como resta, multiplicación y división:

```
> Var1-2  
[1] 1 0 2  
> Var1*4  
[1] 12 8 16  
> Var1/3  
[1] 1.0000000 0.6666667 1.3333333
```

La primera operación restó 2 a los valores 3, 2 y 4 que estaban dentro de la variable “Var1”, la segunda multiplicó 4 y en la tercera dividió entre 3. Los resultados se presentan en la línea de respuesta 1 de cada operación.

También podemos crear un vector adicional para sumarlo al primer vector, por ejemplo, crearemos un vector con los valores 7, 1, 5 y los guardaremos en la variable “Var2”:

```
> Var2 <-c(7,1,5)
```

Ahora sumamos ambos vectores Var1+Var2:

```
> Var1+Var2  
[1] 10 3 9
```

El programa sumó los valores del vector 1 guardados en “Var1” (3, 2 y 4) y los valores del vector 2 guardados en “Var2” (7, 1 y 5), mediante la operación 3+7, 2+1, 4+5 y presentó como resultado 10, 3 y 9. De igual forma podemos realizar otras operaciones como resta, multiplicación y división entre vectores. La premisa principal es que los vectores deben de tener la misma cantidad de números, en el ejemplo Var1 tiene tres números y Var2 tiene tres números, de lo contrario el programa desplegará un mensaje de advertencia. Por ejemplo, si creamos el vector 3 en la variable “Var3” con cuatro caracteres (4,6,2,6) en lugar de tres y corremos la operación Var1+Var3 el resultado es el siguiente:


```
> Var3 <-c(4,6,2,6)
> Var1+Var3
[1] 7 8 6 9
Warning message:
In Var1 + Var3 :
  longer object length is not a multiple of shorter object
  length
```

El mensaje advierte (Warning message) que Var1 y Var3 tienen diferentes longitudes de objetos, o sea diferente cantidad de números.

El razonamiento del vector es parecido al de un conjunto de datos guardados en una columna y varias filas de una hoja de cálculo de MS Excel o en cualquier otro formato tabular; las operaciones se realizan de forma similar, las respuestas son las mismas, lo único que cambia es el cómo se ve la organización de los números. En el formato tabular el nombre de la variable se escribe en el encabezado de cada columna y los números se presentan en las filas que conforman las columnas; en el formato vectorial los números que conforman el vector están guardados en variables y no son visibles, como en el formato tabular (Figura 77).

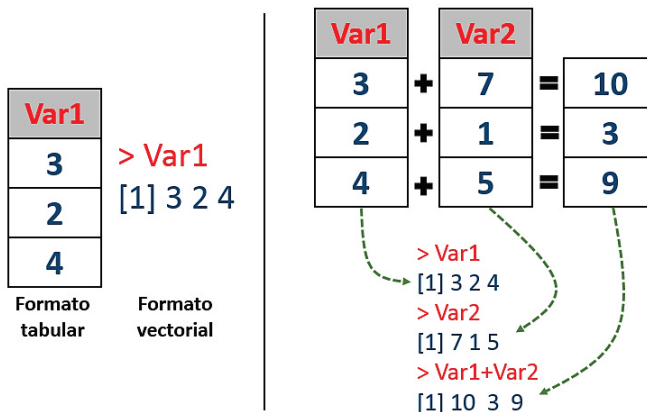


Figura 77. Representación gráfica de las similitudes y diferencias entre el formato tabular y el vectorial. A la izquierda las diferencias entre los datos organizados en formato tabular en MS Excel y los datos organizados en formato vectorial en R; a la derecha una ejemplificación de una operación de suma de los valores en ambos formatos.

Hay diferentes formas para estructurar vectores, las formas manuales donde el operador escribe los vectores (como las ejemplificadas anteriormente) y hay formas automáticas de generar vectores con números que contemplen valores y órdenes lógicos. En el cuadro 8 se presentan y ejemplifican la mayoría de las opciones para generar vectores de forma automática.

Aplicaciones de Estadística Básica

Cuadro 8. Formas comunes de estructurar un vector siguiendo secuencias lógicas.

Operadores/ Funciones	Descripción	Ejemplo
Dos puntos ":"	Crea un vector con números continuos de una cantidad de valores indicados.	<pre>> Vector1 <-1:4 > Vector1 [1] 1 2 3 4</pre>
seq(,by)	Crea un vector con una secuencia dentro de un rango de números.	<pre>> Vector2 <-seq(1,4, by=0.5) > Vector2 [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0</pre>
rep(,times)	Crea un vector con números repetidos en cantidades indicadas.	<pre>> Vector3 <-rep(3:4, times=4) > Vector3 [1] 3 4 3 4 3 4 3 4</pre>
rep(,each)	Crea un vector con una secuencia de repeticiones según número indicado.	<pre>> Vector4 <-rep(2:4, each=3) > Vector4 [1] 2 2 2 3 3 3 4 4 4</pre>

Dentro de los vectores podemos seleccionar elementos según la posición que ocupen dentro de los mismos. Si creamos un vector llamado "X" al que le asignemos los valores 2, 4, 5, 3, 6, 1, 2, 4 y 5, el vector quedaría definido de la siguiente forma:

```
> X <-c(2,4,5,3,6,1,2,4,5)
> X
[1] 2 4 5 3 6 1 2 4 5
```

Notemos que cada número dentro del vector tiene una posición, el número 2 está en la primera posición, el número 4 en la segunda, el número 5 en la tercera y así sucesivamente todos los elementos dentro del vector tienen una posición. Con esa premisa, podemos seleccionar los elementos por ejemplo, si escribimos el nombre de la variable y entre corchetes la posición del objeto en el vector devuelve el valor del objeto en esa posición:

```
> X[5]
[1] 6
```

Le hemos indicado al programa que en el vector guardado en la variable llamada "X" buscara al objeto en la posición 5 y ese objeto fue el valor 6. Con el operador dos puntos (":") podemos llamar a los valores que estén en un rango de posición:

```
> X[3:6]
[1] 5 3 6 1
```

En Microsoft® Excel y R

Le indicamos al programa que seleccione los valores de la posición 3 a la 6 dentro del vector. Así también podemos solicitar los valores en posiciones discontinuas, por ejemplo solicitar los valores de la posición 3 y 8:

```
> X[c(3, 8)]  
[1] 5 4
```

Los valores que están en la posición 3 y 8 del vector fueron 5 y 4 respectivamente. Así como hemos extraído algunos valores del vector ignorando el resto de ellos, así podemos extraer todos los valores ignorando alguno de ellos, esto lo logramos cuando asignamos el valor negativo a las posiciones solicitadas:

```
> X[-5]  
[1] 2 4 5 3 1 2 4 5  
> X[-(3:6)]  
[1] 2 4 2 4 5  
> X[c(-3, -8)]  
[1] 2 4 3 6 1 2 5
```

Notemos que asignando el valor negativo a 5 el programa presenta todos los valores del vector, excepto el que está en la posición 5, el cual es el número 6; la otra operación presenta todos los valores excepto los que están en el rango de posición de la 3 a la 6; y el último ejemplo presenta todos los valores excepto los que están en la posición 3 y 8. Los vectores no solamente están formados por números, también pueden estar formados por palabras y otros objetos u elementos. Por ejemplo, formaremos un vector con cinco objetos que corresponden a nombre de personas (Martha, María, Oscar, Rodrigo y Luisa), la diferencia con los números es que los vectores de objetos cualitativos se guardan poniendo comilla a los mismos:

```
> Nombres <-c("Martha", "María", "Oscar", "Rodrigo", "Luisa")  
> Nombres  
[1] "Martha" "María" "Oscar" "Rodrigo" "Luisa"
```

Como vemos, el vector guardado en la variable “Nombres” ahora está formado por elementos no numéricos.

El formato tabular

Aprender a manejar vectores en R es muy importante; sin embargo, la mayor parte de los usuarios prefieren el trabajo en formato tabular (o matricial), los cuales están conformados por filas y columnas. R presta todas las condiciones para trabajar de esta ma-

Aplicaciones de Estadística Básica

nera también. A continuación se mostrarán dos formas de cómo se originan los datos en formato tabular en R; uno con la conversión de un conjunto de datos de un formato vectorial a otro tabular y otro mediante la importación de datos con formatos tabulares almacenados en otros programas (como MS Excel).

A partir de vectores, podemos generar datos en formato tabular utilizando las correspondientes funciones de transformación. A continuación, vamos a crear tres vectores relacionados y a partir de ellos vamos a crear un pequeño conjunto de datos en forma tabular. El primer vector será una cuenta consecutiva de seis árboles (del 1 al 6), la cual almacenaremos en la variable “Árbol”; el segundo vector será la localización donde se encuentran los seis árboles, los primeros tres se encuentran en un bosque secundario (BS) y el resto en bosque primario (BP), esta información la incluiremos en la variable “Sitio” y finalmente el tercer vector estará formado por las medidas del diámetro a la altura del pecho (DAP) de los seis árboles, la guardaremos en la variable “DAP”. En R, la creación de los tres vectores se vería de la siguiente manera:

```
> Arbol <-c(1,2,3,4,5,6)
> Sitio <-c("BS","BS","BS","BP","BP","BP")
> DAP <-c(37.2,23.1,12.3,59.6,89.6,65.9)
```

Lo que queremos hacer es combinar esos tres vectores, para formar un pequeño conjunto de datos en formato tabular, para esto vamos a hacer uso de la función “data.frame()”, la cual tomará los tres vectores y los combinará de una forma lógica, en la cual los primeros tres elementos de los tres vectores se pondrán juntos en una sola fila, los segundos elementos de los tres vectores se pondrán juntos en la segunda fila y así sucesivamente. El orden en que escribamos los vectores será el orden en las columnas de la tabla de datos, la cual a la vez, la guardaremos en una variable que nombraremos “DatoTabular”:

```
> DatoTabular <-data.frame(Arbol,Sitio,DAP)
> DatoTabular
  Arbol Sitio  DAP
1     1   BS 37.2
2     2   BS 23.1
3     3   BS 12.3
4     4   BP 59.6
5     5   BP 89.6
6     6   BP 65.9
```

Notemos que la función “data.frame()” establece el nombre del vector como nombre de las columnas, estos nombres los podemos personalizar dentro de la misma función y antes de formar, haremos esta asignación y guardaremos para ilustrar escribiendo Vector1, Vector2 y Vector3 a las columnas, en lugar de los nombres, los resultados los guardaremos en otra variable a la que llamaremos “DatoTabular2”:

```
> DatoTabular2 <-data.frame(Vector1=Arbol, Vector2=Sitio,
Vector3=DAP)
> DatoTabular2
  Vector1 Vector2 Vector3
1        1      BS   37.2
2        2      BS   23.1
3        3      BS   12.3
4        4      BP   59.6
5        5      BP   89.6
6        6      BP   65.9
```

Al igual que en formato vectorial, en el formato tabular también podemos seleccionar datos sobre la base de la posición, formadas por la combinación de filas y columnas, para ello utilizaremos los corchetes y los números que indican las filas y las columnas separados por una coma. En general el primer número representa a las filas y el segundo a las columnas o sea: [filas,columnas], de tal forma que la expresión [,3] está indicando la columna 3, la expresión [1,] está indicando la fila 1; la expresión [1,3] indica el valor que se encuentra en la fila 1 y columna 3:

```
> DatoTabular[,3]
[1] 37.2 23.1 12.3 59.6 89.6 65.9
> DatoTabular[1,]
  Arbol Sitio DAP
1      1    BS 37.2
> DatoTabular[1,3]
[1] 37.2
```

De los datos en formato tabular extrajimos la información de la columna 3, de la fila 1 y el dato en la posición fila 1 columna 3. Notar que la información extraída se presenta en formato vectorial, esto tiene implicaciones en el uso de dicha información. Como en el formato vectorial, en el formato tabular podemos hacer algunas combinaciones de llamados, por ejemplo extraer columnas, filas o valores en particular, a como se reflejan en los ejemplos del cuadro 9.

Aplicaciones de Estadística Básica

Cuadro 9. Ejemplificaciones para la extracción o llamado de datos de filas y columnas dentro de una tabla de datos.

<p>Se extrajeron los datos de las primeras cuatro filas:</p> <pre>> DatoTabular[1:4,] Arbol Sitio DAP 1 1 BS 37.2 2 2 BS 23.1 3 3 BS 12.3 4 4 BP 59.6</pre>	<p>Se extrajeron la columna 1 (Arbol) y 2 (Sitio):</p> <pre>> DatoTabular[,1:2] Arbol Sitio 1 1 BS 2 2 BS 3 3 BS 4 4 BP 5 5 BP 6 6 BP</pre>
<p>Se extrajeron las columnas 1 (Arbol) y 3 (DAP):</p> <pre>> DatoTabular[,c(1,3)] Arbol DAP 1 1 37.2 2 2 23.1 3 3 12.3 4 4 59.6 5 5 89.6 6 6 65.9</pre>	<p>Se extrajeron las filas 1 y 4, y las columnas 1 (Arbol) y 3 (DAP).</p> <pre>> DatoTabular[c(1,4),c(1,3)] Arbol DAP 1 1 37.2 4 4 59.6</pre>

Asignando el signo negativo, podemos extraer los datos, excluyendo algún subconjunto de información:

```
> DatoTabular[-(5:6),c(1,2)]
  Arbol Sitio
1     1    BS
2     2    BS
3     3    BS
4     4    BP
```

El argumento realizó la extracción de los datos que están en las columnas 1 y 2, exceptuando la información que está en las filas de la 5 a la 6.

Con el uso de los encabezados de las variables, en la tabla de datos, también podemos hacer llamado de valores, utilizando los operadores de igualdad (==), mayor (>), mayor e igual (>=), menor (<), menor e igual (<=) (Cuadro 10).

Cuadro 10. Ejemplificación de selecciones basadas en operadores de igualdad o desigualdad. En el ejemplo se seleccionan filas en base a criterios asignados a ciertas columnas.

<p>Se extrajo la fila (registro completo) a partir del dato igual a 59.6 en la columna DAP:</p> <pre>> DatoTabular[DAP==59.6,] Arbol Sitio DAP 4 4 BP 59.6</pre>	<p>Se extrajeron todas las filas (registros completos) que tenía en la variable DAP valores menores e igual a 37.2:</p> <pre>> DatoTabular[DAP<=37.2,] Arbol Sitio DAP 1 1 BS 37.2 2 2 BS 23.1 3 3 BS 12.3</pre>
<p>Se extrajeron todas las filas (registros completos) que tenía en la variable DAP valores mayores e igual a 37.2:</p> <pre>> DatoTabular[DAP>=37.2,] Arbol Sitio DAP 1 1 BS 37.2 4 4 BP 59.6 5 5 BP 89.6 6 6 BP 65.9</pre>	<p>Se extrajeron todas las filas (registros completos) que tenía en la variable "Sitio" la categoría "BS":</p> <pre>> DatoTabular[Sitio=="BS",] Arbol Sitio DAP 1 1 BS 37.2 2 2 BS 23.1 3 3 BS 12.3</pre>

Adicionalmente, podemos extraer datos con la función "subset()" incluyendo como argumento el nombre de la variable donde se encuentran los datos y la selección. Por ejemplo, de los datos que hemos guardado en la variable "DatoTabular" vamos a extraer con esta función las columnas "Arbol" y "DAP" y los resultados los guardaremos en una variable llamada "Arbol_DAP". Notemos que dentro del argumento "select=" los nombres de las columnas no se escriben entre comillas:

```
> Arbol_DAP <-subset(DatoTabular, select=c(Arbol,DAP))  
> Arbol_DAP  
  Arbol  DAP  
1      1 37.2  
2      2 23.1  
3      3 12.3  
4      4 59.6  
5      5 89.6  
6      6 65.9
```

Aplicaciones de Estadística Básica

Ahora vamos a extraer las filas de datos, que correspondan solamente con la categoría “BP” de la columna “Sitio”, utilizando el argumento “subset=” dentro de la función del mismo nombre y la guardaremos en la variable Solo_BP. Notemos que la selección la realizamos escribiendo el nombre de la columna, doble igual y el nombre de la categoría a seleccionar:

```
> Solo_BP <-subset(DatoTabular, subset=(Sitio=="BP"))
> Solo_BP
  Arbol Sitio  DAP
4      4     BP 59.6
5      5     BP 89.6
6      6     BP 65.9
```

También podemos seleccionar rangos de número con sus respectivos operadores, por ejemplo si quisiéramos seleccionar los valores de la variable numérica “DAP”, que sean mayor de 60, el comando quedaría estructurado de la siguiente manera:

```
> Mayores60 <-subset(DatoTabular, subset=(DAP>=60))
> Mayores60
  Arbol Sitio  DAP
5      5     BP 89.6
6      6     BP 65.9
```

Ahora seleccionaremos valores de “DAP” entre 30 y 60 agregando la apertura y el cierre del rango de número, separado por el simbolo ampersand “&”:

```
> Entre30y60 <-subset(DatoTabular, subset=(DAP>=30 & DAP<=60))
> Entre30y60
  Arbol Sitio  DAP
1      1     BS 37.2
4      4     BP 59.6
```

Notemos que para seleccionar las columnas utilizamos el argumento “select=” y para seleccionar las filas utilizamos el argumento “subset=”. Así podemos seleccionar columnas y filas. Por ejemplo, podemos seleccionar las columnas “Arbol” y “DAP” y dentro de esa selección seleccionar los árboles mayores a 60:

```
> DobleSeleccion <-subset(DatoTabular, select=c(Arbol,DAP),
subset=(DAP>=60))
> DobleSeleccion
  Arbol  DAP
5      5 89.6
6      6 65.9
```


Incluso podemos seleccionar a la vez, las columnas “Sitio” y “DAP”, luego los sitios “BP” y dentro de esa selección seleccionar los árboles mayores a 60:

```
> TripleSeleccion <-subset(DatoTabular, select=c(Sitio,DAP),
subset=(Sitio=="BP" & DAP>=60))
> TripleSeleccion
  Sitio  DAP
5    BP 89.6
6    BP 65.9
```

La extracción de datos es de suma importancia cuando trabajamos con bases de datos, pues la mayoría de veces no es necesario trabajar con toda la base, sino con parte de la misma. Además de las antes vista, R tiene una serie de funciones y argumentos dedicados a esta tarea, sin embargo administrar datos en R se hace un tanto complicado en especial cuando se trata de bases de datos de grandes dimensiones. Una opción sencilla es simplemente utilizar otro programa para administrar las bases de datos, como MS Excel, y solamente importar a R las secciones de las bases de datos con las que se van a hacer análisis o gráficos.

Para importar tablas de datos en R debemos seguir algunos pasos: Si tenemos el archivo en MS Excel, primero lo tenemos que guardar en formato “delimitado por tabulaciones” o “delimitado por coma”, en particular para este ejemplo lo guardaremos con esta última opción, seleccionando (estando en MS Excel) la secuencia de opciones Archivo>-Guardar>CSV(delimitado por comas)(Paso 1) (Figura 78); aparecerá una advertencia que dirá “Algunas características de su libro pueden perderse si lo guarda como CSV (delimitado por comas)” y preguntará “¿Desea seguir utilizando este formato?”, para lo cual seleccionamos “Sí” (Paso 2). El archivo se llama “Datos_Muestra” y es importante saber exactamente dónde se guarda el archivo, pues esa ubicación será utilizada en los siguientes pasos.

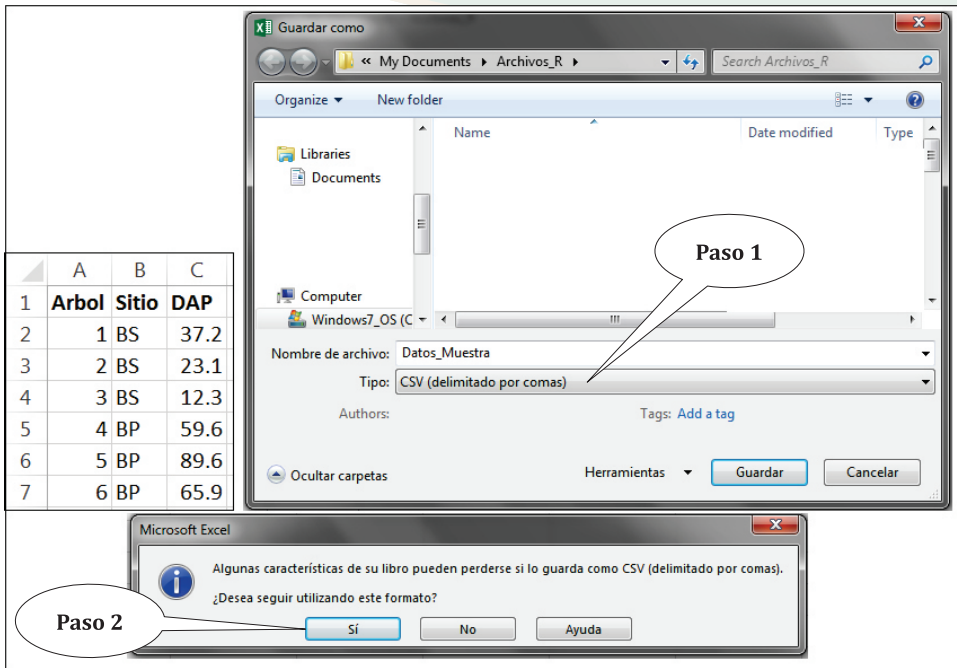


Figura 78. Pasos para guardar en formato CSV (texto delimitado por comas) una tabla o base de datos que está en una hoja de cálculo de MS Excel.

Seguido abrimos el programa R y en la consola primero vamos a establecer el directorio de trabajo, o simplemente, la carpeta donde está guardado el archivo con el que se pretende trabajar. Esto lo realizamos con la función “setwd()” y dentro del paréntesis debemos escribir la dirección de la carpeta donde el archivo está guardado. Una forma sencilla de obtener esta dirección es haciendo clic derecho sobre el archivo (en el explorador) que se pretende abrir en R, seleccionando la opción de “Propiedades” y dirigiéndonos a la información que aparece en la dirección de “Ubicación” (pestaña “General”) (Paso 3) (Figura 79); esa dirección la debemos copiar y pegar dentro de los paréntesis de la función “setwd()”, la ponemos entre comillas y cambiamos las plecas hacia la izquierda por plecas hacia la derecha. La dirección en el recuadro de “Propiedades” es: C:\Users\Miguel Garmendia\Documents\Archivos_R, la dirección ingresada en la función se tendría que ver de esta forma: setwd(“C:/Users/Miguel Garmendia/Documents/Archivos_R”).

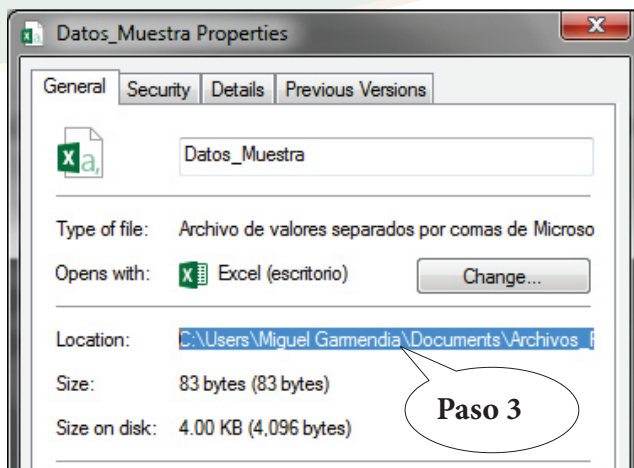


Figura 79. Extracción de la ubicación en la cual se encuentra el archivo guardado en la computadora del usuario, haciendo uso de la opción de “Propiedades” del archivo.

Seguidamente podemos utilizar la función “dir()” para enlistar los archivos que se encuentran en la dirección recientemente establecida:

```
> setwd("C:/Users/Miguel Garmendia/Documents/Archivos_R")
> dir()
[1] "Datos_Muestra.csv"          "DATOS_PORELLA.csv"
[3] "OnewayANOVARepetated.csv"   "RadiacSol.csv"
[5] "RadiacSol2.csv"             "RadiacSol3.csv"
```

En nuestro ejemplo, es necesario importar a R el archivo llamado “Datos_Muestra.csv”, para ello hacemos uso de la función “read.csv()” y dentro del paréntesis escribimos el nombre del archivo (nombre exacto) con su extensión y entre comillas, y los guardamos en una variable que llamaremos “DatosTabular”:

```
> DatoTabular <- read.csv("Datos_Muestra.csv")
> DatoTabular
  Arbol Sitio DAP
1     1    BS 37.2
2     2    BS 23.1
3     3    BS 12.3
4     4    BP 59.6
5     5    BP 89.6
6     6    BP 65.9
```

Aplicaciones de Estadística Básica

Con la primera línea de comando orientamos al programa a que importe los datos y los guarde en la variable llamada “DatoTabular”, la tabla de datos ya quedó lista para realizar análisis con ella. Notemos que esta tabla es idéntica a la que elaboramos con el uso de vectores. Si hubiéramos grabado la tabla de dato en formato TXT (texto delimitado por tabulaciones) la función y el argumento (comando) para ingresar los datos en R hubiese quedado escrita como: `read.table(“Datos_Muestra.txt”)`.

Otra opción, y la más sencilla, es el utilizar la función “`file.choose()`” dentro de la función “`read.csv()`”, la cual nos abre la ventana del explorador y simplemente podemos buscar el archivo y asignarlo directamente a una variable (Figura 80):

```
> DatoTabular <-read.csv(file.choose())
```

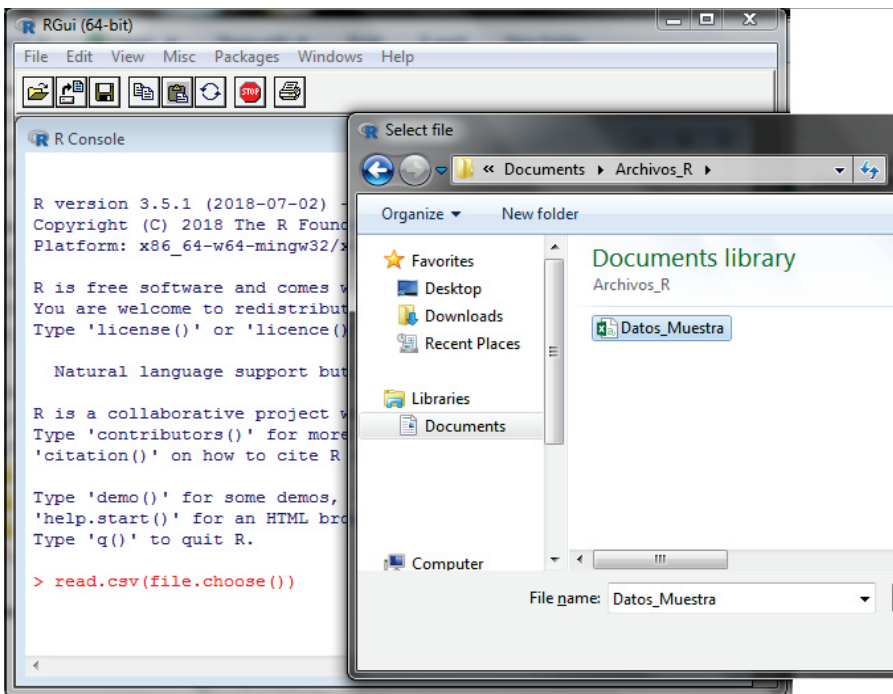


Figura 80. Ilustración de una forma de importar a R datos tabulares grabados en formato CSV, utilizando la función “`file.choose()`”. Al hacer Enter al comando “`DatoTabular <-read.csv(file.choose())`” se despliega una ventana de exploración donde se puede seleccionar el archivo.

La información recién importada con el uso de la función “`file.choose()`” muestra la misma información que la que muestran las formas explicadas anteriormente.

```
> DatoTabular
  Arbol Sitio  DAP
1     1     BS 37.2
2     2     BS 23.1
3     3     BS 12.3
4     4     BP 59.6
5     5     BP 89.6
6     6     BP 65.9
```

Además de esta función de lectura (“read.csv”), R cuenta con una función de escritura con la que podemos guardar tablas en el directorio que estamos usando en la computadora, esta función es “write.csv()”. Por ejemplo, si quisiéramos guardar la tabla de datos guardada en la variable “DatoTabular” en la computadora, tendríamos que usar el comando:

```
> write.csv(DatoTabular, file="DatosDAP.csv", col.names=TRUE)
```

Dentro de la función “write.csv()” hay tres argumentos, el primero es la variable donde están guardados los datos; el segundo es la asignación del nombre que tendrá el archivo guardado (DatosDAP) más la extensión CSV (Texto Delimitado por Comas); el tercero le indica al programa que la tabla tiene encabezados y debe mantenerlos.

Con las siguientes funciones podemos hacer conversiones de cualquier formato a formatos vectoriales, tabulares y matriciales:

- La función “as.data.frame()” transforma a formato tabular.
- La función “as.matrix()” transforma a formato matricial.
- La función “as.vector()” transforma a formato vectorial.
- También existe la función “as.list()” para generar una lista.

Al trabajar con tablas de datos, una actividad cotidiana es la transposición, es decir que los datos ordenados en columnas pasen a estar ordenados por fila y que los datos ordenados por fila pasen a estar ordenados por columnas. Esto lo logramos fácilmente con la función “t()” y como argumento asignamos la tabla de datos a transponer. Por ejemplo, si deseamos transponer la tabla de datos guardada en la variable “DatoTabular”, el comando y el resultado quedaría escrito a como se presenta seguidamente, el cual lo guardaremos en la variable “DatoTabularT”:

Aplicaciones de Estadística Básica

```
> DatoTabularT <-t(DatoTabular)
> DatoTabularT
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
Arbol	"1"	"2"	"3"	"4"	"5"	"6"
Sitio	"BS"	"BS"	"BS"	"BP"	"BP"	"BP"
DAP	"37.2"	"23.1"	"12.3"	"59.6"	"89.6"	"65.9"

La tabla resultante realmente se encuentra en formato matricial, esto lo notamos porque los elementos que contiene se presentan entre comillas. Los nombres de las columnas ahora son los registros de las filas, para cambiar los nombres de las columnas, tendríamos que utilizar la función “colnames()” y para cambiar los nombres de las filas utilizaremos la función “rownames()”. Ejemplificaremos la asignación de nombres a las columnas, además cómo transformar la tabla de datos en formato tabular con la función “as.data.frame()” y guardaremos los resultados en una nueva variable llamada “DatoTabularT2”:

```
> colnames(DatoTabularT) <-c(1:6)
> DatoTabularT2 <-as.data.frame(DatoTabularT)
> DatoTabularT2
```

	1	2	3	4	5	6
Arbol	1	2	3	4	5	6
Sitio	BS	BS	BS	BP	BP	BP
DAP	37.2	23.1	12.3	59.6	89.6	65.9

Notemos que las columnas ahora tienen encabezados (números del 1 al 6) y los datos se presentan en formato tabular. Si necesitamos saber el formato en el que se encuentran nuestros datos podemos utilizar la función “class()”:

```
> class(DatoTabular)
[1] "data.frame"
> class(DatoTabularT)
[1] "matrix"
```

Mediante la función “class()”, R nos está indicando que los datos guardados en la variable “DatoTabular” tienen formato tabular (data.frame) y los datos guardados en la variable “DatoTabularT” tienen formato de matriz (matrix).

Instalación de paquetes

Los paquetes son complementos del programa R, los cuales contienen funciones que no se encuentran disponibles en R básico. Los paquetes son creados por toda una comunidad internacional; sin embargo, no hay garantía de su utilización, la única garantía que se tiene es la confianza que la comunidad internacional deposita en los colaboradores que los elaboran, basada en la buena reputación de los mismos.

En sí, hay tres formas de instalar un paquete, la primera forma la utilizaremos cuando se ha instalado solamente el programa R, sin algún entorno secundario como “RStudio” (un entorno de desarrollo integrado o IDE para el lenguaje de programación R) haciendo uso de la opción “Install package(s)...” que se encuentra en pestaña “Packages” del programa (Paso 1) (Figura 81). Para ejecutar la opción tenemos que especificar el “CRAN mirror”, las siglas “CRAN” obedecen a las iniciales de “The Comprehensive R Archive Network” o “La red completa de archivos de R” y “mirror” se traduce como “espejo” en español.

El “CRAN mirror” es una localidad virtual de donde se descargan los paquetes. Como preferencia personal, suelo descargar los paquetes del CRAN mirror “USA(NY)” (Paso 2); sin embargo, el lector es libre de seleccionar la localidad de donde descargar los paquetes y a la vez de revisar más información sobre el tema. Como regla general se suele seleccionar el CRAN mirror más cercano al país en donde se está, pero no es obligatorio. Después de haber seleccionado el CRAN mirror, se selecciona el paquete de interés, para ejemplificar se seleccionará el paquete estadístico llamado “e1071” (Paso 3).

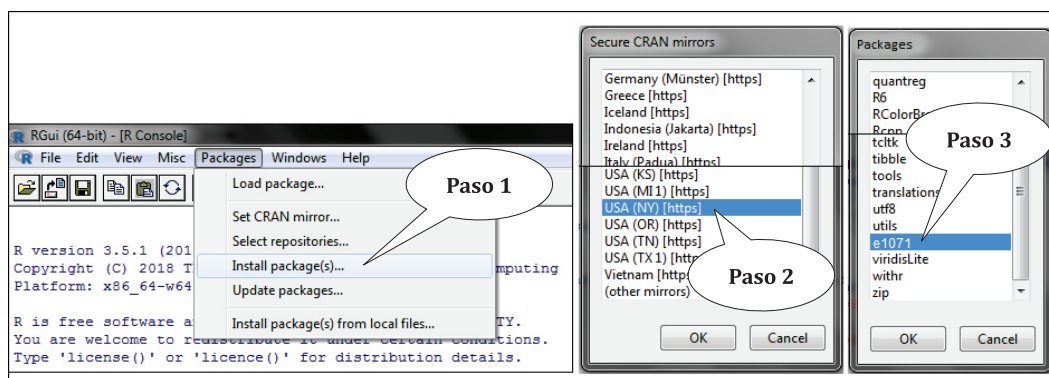


Figura 81. Pasos para la instalación de paquetes en el programa R. Para ejemplificar se instala el paquete “e1071” del CRAN mirror “USA (NY)”.

Aplicaciones de Estadística Básica

Al instalar el paquete “e1071” del CRAN mirror “USA (NY)”, la consola de R presentará información sobre dicho paquete y sobre el proceso de instalación:

```
> utils:::menuInstallPkgs()  
--- Please select a CRAN mirror for use in this session ---  
trying URL 'https://mirrors.sorengard.com/cran/bin/windows/  
contrib/3.2/e1071_1.6-8.zip'  
Content type 'application/zip' length 800519 bytes (781 KB)  
downloaded 781 KB  
  
package 'e1071' successfully unpacked and MD5 sums checked  
  
The downloaded binary packages are in  
  C:\Users\Miguel Garmendia\AppData\Local\Temp\RtmpovWy-  
ut\downloaded_packages  
>
```

Después de finalizada la instalación del paquete es necesario que se cargue, para ello se utiliza la función “library()”:

```
> library("e1071")
```

Hasta este punto el paquete está cargado y listo para funcionar, este proceso de instalación solo lo necesitamos realizar una vez. Para efectuar la instalación, es necesario que la computadora tenga acceso a Internet.

La segunda forma de instalar paquetes funciona con el programa R normal o “RStudio” y para ello utilizaremos la función “install.packages()” y entre el paréntesis especificaremos el nombre del paquete entre comillas, ej.: `install.packages("e1071")`. Luego de instalado, el paquete lo haremos activos con la función “library()”, a como se especificó anteriormente.

La tercera forma de instalar paquetes es descargándolos directamente del sitio Web: https://cran.r-project.org/web/packages/available_packages_by_name.html, este presentará la lista por orden alfabético de todos los paquetes, simplemente seleccionamos el deseado, extraemos el instalador que está compreso y procedemos a instalarlo manualmente.

Además de la función “library()” para cargar un programa y hacerlo disponible, R cuenta con otra opción para hacer la misma tarea, siguiendo la secuencia de opciones `Packages>Load packages...` y seleccionamos el programa, la misma es útil también para visualizar todos los paquetes ya instalados en el ordenador.

Sobre el entorno de RStudio

El entorno de desarrollo integrado (IDE) “RStudio” ofrece una opción visualmente más confortable, versátil y más amigable con el usuario para trabajar con el programa R. Este funciona solamente después de haber instalado R y se puede descargar del sitio Web: <https://www.rstudio.com/products/rstudio/download/>. RStudio utiliza al mismo programa R para realizar las operaciones, pero el ambiente de trabajo es muy diferente, aunque las funciones y codificaciones son las mismas. A diferencia del ambiente del programa R meramente dicho, RStudio está conformado por cuatro ventanas de trabajo, estas son: la ventana “R Script”, la ventana “Console”, la ventana “Environment – History – Connections” y la ventana “Files – Plots – Packages – Help – Viwers” (Figura 82).

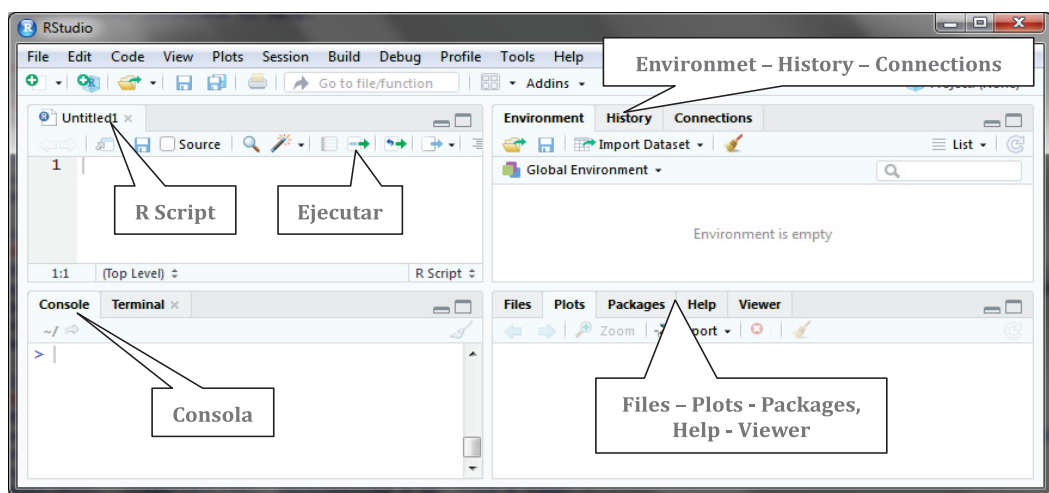


Figura 82. Ilustración de las partes del entorno RStudio para el programa R.

En la ventana “R Script” escribimos los comandos a ejecutar mediante el sistema de codificación de R, el R Script presenta varias opciones para facilitar la escritura: anticipando palabras, funciones o argumentos; poniendo los dos paréntesis, corchetes o comillas al poner solamente el paréntesis, corchete o comilla de apertura; entre otras facilidades, lo cual se traduce en una invaluable cantidad de tiempo ganado. En esta ventana la tecla Enter no ejecuta acciones, sino agrega una línea de espacio, para que ejecute los comando necesitamos hacer clic en “Run” (“correr” en español) o pulsar las teclas “Ctrl + Enter”.

Aplicaciones de Estadística Básica

En la ventana “Console” se despliegan los comandos y las respuestas de la ejecución de esos comandos, es la misma consola que se ocupa en el programa R meramente dicho; incluso, se pueden escribir directamente los códigos en esta, si se considera no utilizar la ventana “R Script”, pero en la ventana “Console” no tenemos facilidades de escritura. En la ventana “Environment – History” observamos la estructura de los datos y el historial de trabajo en RStudio. Y en la ventana “Files – Plots – Packages – Help – Viwers”, tenemos las opciones de: dirigirnos al explorador en la pestaña “Files”; visualizar los gráficos en la pestaña “Plots”; explorar la lista de paquetes instalados en la pestaña “Packages”; acceder al menú de ayuda de R y ver todos los objetos de trabajo en la pestaña “Viwers”.

En este documento se está haciendo énfasis al trabajo directamente con el programa R con el entorno básico (Console), no se hará referencia al entorno RStudio; sin embargo, queda a decisión del lector el explorar el uso de RStudio en su trabajo cotidiano.

Estadística descriptiva en R

Aplicando estadística descriptiva

La estadística descriptiva, como su nombre lo indica, se utiliza para describir y explorar los datos en términos de su tendencia central, la dispersión y la forma, ello permite calcular medidas como media, mediana, desviación estándar, error estándar, curtosis, entre otras. Para ejemplificar, ingresaremos algunos datos en formato vectorial y posteriormente lo trasformaremos en formato tabular a como se explicó en el capítulo anterior, seguidamente a la columna de datos aplicaremos estadística descriptiva.

Utilizaremos un pequeño conjunto de datos de mediciones de potencial de hidrógeno (pH) del suelo en cinco puntos diferentes, en un área agrícola, los datos en R se verían de la siguiente forma:

```
> pH <-c(4.7,5.3,5.9,4.9,4.6)
> PUNTO <-(1:5)
> Info_pH <-data.frame(PUNTO,pH)
> Info_pH
  PUNTO  pH
1     1 4.7
2     2 5.3
3     3 5.9
4     4 4.9
5     5 4.6
```

En la primera línea de comando creamos el vector que contiene los cinco valores de pH, en la segunda línea creamos otro vector que contiene número consecutivo del 1 al 5 que representará cada punto donde se tomó el pH en el suelo; en la tercera línea fusionamos los dos vectores para formar una tabla de datos (formato tabular) y la información resultante la guardamos en la variable “Info_pH” y en la cuarta línea desplegamos la información guardada en la variable “Info_pH”, escribiendo el nombre de la variable y presionando Enter.

Aplicaremos estadística descriptiva a los datos de pH, utilizando la función “summary()”, para ello tendríamos que indicarle al programa la columna de datos, a la cual se le aplicará el análisis, esto lo logramos asignándole a la función el nombre de la variable y el encabezado de la columna separados por el signo de dólar (\$), o sea Info_pH\$pH, dentro de la función se vería como: summary(Info_pH\$pH) (Figura 83).

> Info_pH					
	PUNTO	pH			
1	1	4.7			
2	2	5.3			
3	3	5.9			
4	4	4.9			
5	5	4.6			
(Info_pH\$pH)					

Info_pH		
	PUNTO	pH
1	1	4.7
2	2	5.3
3	3	5.9
4	4	4.9
5	5	4.6

Figura 83. Equivalencias entre el formato tabular utilizado por R y el formato tabular utilizado por las hojas de cálculo de otros programas, incluyendo MS Excel.

La ejecución de la función “summary()” para los datos de pH de la variable Info_pH se realizaría en R de la siguiente manera:

```
> summary(Info_pH$pH)
  Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
  4.60    4.70    4.90    5.08    5.30    5.90
```

La función ha extraído y calculado simultáneamente el valor mínimo (Min.), el primer cuartil (1st Qu.), la mediana (Median), el promedio (Mean), el tercer cuartil (3rd Qu.) y el valor máximo (Max.). Si el lector no está familiarizado con alguno de los términos, le recomendamos visitar literatura sobre estadística básica.

Podemos utilizar las funciones por separado, e incluso separadamente podemos encontrar más funciones de estadísticas descriptivas, como las que se describen en el cuadro 11.

Aplicaciones de Estadística Básica

Cuadro 11. Funciones para calcular parámetros descriptivos para conjuntos de datos en R.

Función	Descripción	Ejemplo
mean()	Calcula el promedio.	<pre>> mean(Info_pH\$pH) [1] 5.08</pre>
median()	Calcula la mediana.	<pre>> median(Info_pH\$pH) [1] 4.9</pre>
mode()	Calcula la moda. Devuelve “numeric” si no encuentra datos repetidos para calcular moda.	<pre>> mode(Info_pH\$pH) [1] "numeric"</pre>
max()	Presenta el valor máximo.	<pre>> max(Info_pH\$pH) [1] 5.9</pre>
min()	Presenta el valor mínimo.	<pre>> min(Info_pH\$pH) [1] 4.6</pre>
length()	Devuelve la cuenta de objetos.	<pre>> length(Info_pH\$pH) [1] 5</pre>
sd()	Calcula la desviación estándar.	<pre>> sd(Info_pH\$pH) [1] 0.5310367</pre>
range()	Calcula el rango.	<pre>> range(Info_pH\$pH) [1] 4.6 5.9</pre>
var()	Calcula la varianza.	<pre>> var(Info_pH\$pH) [1] 0.282</pre>
quartile()	Calcula los cinco cuartiles.	<pre>> quantile(Info_pH\$pH) 0% 25% 50% 75% 100% 4.6 4.7 4.9 5.3 5.9</pre>

A continuación agregamos una nueva columna a la tabla de datos guardada como “Info_pH”, a esta nueva columna le llamaremos “HS” y tendrá datos de humedad del suelo en porcentaje. Adicionalmente, cambiaremos el número que designa a cada punto por letras, solamente para ejemplificar el uso de vectores con letras en lugar de números. La nueva tabla de datos la guardaremos con el nombre de “Info_pH_HS”:

```
> pH <-c(4.7, 5.3, 5.9, 4.9, 4.6)
> PUNTO <-c("A", "B", "C", "D", "E")
> HS <-c(83.2, 95.4, 74.3, 87.6, 69.7)
> Info_pH_HS <-data.frame(PUNTO, pH, HS)
> Info_pH_HS
  PUNTO  pH  HS
1     A 4.7 83.2
2     B 5.3 95.4
3     C 5.9 74.3
4     D 4.9 87.6
5     E 4.6 69.7
```

La nueva tabla contenida en la variable “Info_pH_HR” con tres columnas “PUNTO”, “pH” y “HS” será utilizada para ejemplificar cálculos de valores descriptivos a tablas con varias columnas. En principio la función “summary()” puede calcular estadísticas de toda la tabla completa:

```
> summary(Info_pH_HS)
PUNTO      pH      HS
A:1  Min.   :4.60  Min.   :69.70
B:1  1st Qu.:4.70  1st Qu.:74.30
C:1  Median :4.90  Median :83.20
D:1  Mean   :5.08  Mean   :82.04
E:1  3rd Qu.:5.30  3rd Qu.:87.60
     Max.   :5.90  Max.   :95.40
```

El programa aplicó estadística descriptiva a toda la tabla de datos guardada en la variable “Info_pH_HS”. Notar que para la columna “PUNTO” como contiene datos categóricos (A, B, C, etc.), no aplicó los mismos cálculos en relación a las columnas con datos numéricos (pH y HS), sino solamente presentó la cuenta, para la cual cada letra solamente está repetida una vez.

R ofrece una forma para poder seleccionar las columnas, a las cuales se precisa aplicarles la función “summary()”, esto lo logramos utilizando los corchetes y la función de combinar “c()”: dentro de los corchetes definimos las filas y las columnas a como se ejemplifica en el cuadro 9 y cuadro 10 ; con la función “c()” especificamos los nombres de las columnas:

```
> summary(Info_pH_HS[,c("pH", "HS")])
      pH      HS
Min.   :4.60  Min.   :69.70
1st Qu.:4.70  1st Qu.:74.30
Median :4.90  Median :83.20
Mean   :5.08  Mean   :82.04
3rd Qu.:5.30  3rd Qu.:87.60
Max.   :5.90  Max.   :95.40
```

Aplicaciones de Estadística Básica

Cuadro 12. Ejemplificación del uso de los corchetes para designar filas y columnas en una tabla de datos en R. Notar la función de la coma entre los corchetes para la designación de las filas y las columnas.

<code>[fila, columna]</code>	Dentro del corchete la coma separa las filas a su izquierda y las columnas a su derecha.
<code>[2, 3]</code>	Llamado al dato que está en la fila 2 y la columna 3.
<code>[2,]</code>	Llamado a los datos que están en la fila 2.
<code>[, 3]</code>	Llamado a los datos que están en la columna 3.
<code>[1:4,]</code>	Llamado a los datos que están de la fila 1 a la fila 4 (incluyendo la fila 2 y 3).
<code>[c(1, 4),]</code>	Llamado a los datos que están solamente en las filas 1 y la 4 (sin incluir ninguna otra fila).
<code>[c(1, 4), c("pH", "HS")]</code>	Llamado a los datos que están solamente en las filas 1 y la 4 y en las columnas con encabezado "pH" y "HS".

Si en lugar de tener una tabla de datos donde las variables categóricas se muestran en columnas separadas (arreglo por columnas) están arregladas en tablas de datos en la que las variables categóricas están apiladas en una sola columna (arreglo por filas) (Figura 84), la función "summary()" la podemos utilizar para cada nivel de la variable categórica con la función "tapply()". Para ejemplificarlo, crearemos una nueva tabla de datos con el arreglo por filas utilizando unos datos de humedad del suelo tomado en dos sitios, a los que llamaremos "Sitio1" y "Sitio2", seguidamente unimos las dos variables y guardamos el resultado en otra variable llamada "HS2", para ello usamos la función "stack()" con el argumento "list()" y definimos el nombre que se repite para cada dato, por ejemplo "Sitio1=Sitio1" significa que para cada valor que guardamos en la variable "Sitio1" asignamos el nombre de "Sitio1" en una columna aparte:

```
> Sitio1 <-c(85.4, 91.2, 93.4, 84.3, 86.5)
> Sitio2 <-c(81.2, 72.1, 82.3, 83.4, 90.1)
> HS2 <-stack(list(Sitio1=Sitio1, Sitio2=Sitio2))
> HS2
  values      ind
1  85.4 Sitio1
2  91.2 Sitio1
3  93.4 Sitio1
4  84.3 Sitio1
5  86.5 Sitio1
```

```
6      81.2 Sitio2
7      72.1 Sitio2
8      82.3 Sitio2
9      83.4 Sitio2
10     90.1 Sitio2
```

Notamos inmediatamente que la tabla con arreglo por filas se ha creado exitosamente, sin embargo ha puesto nombres por defecto a cada columna (values y ind) los cuales tendremos que sustituir por nombres más personalizados. Esto lo logramos con la función "colnames()" y la función "c()" arreglados de tal forma como a continuación se presenta:

```
> colnames(HS2) <-c("HS","Sitios")
> HS2
      HS Sitios
1  85.4 Sitio1
2  91.2 Sitio1
...
9   83.4 Sitio2
10  90.1 Sitio2
```

Después que creamos el conjunto de datos para aplicar la función "summary()", con la función "tapply()" procedemos a correr todo el comando estructurándolo de la siguiente forma:

```
> tapply(HS2$HS, HS2$Sitios, FUN=summary)

$Sitio1
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 84.30  85.40   86.50   88.16  91.20   93.40

$Sitio2
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 72.10  81.20   82.30   81.82  83.40   90.10
```

Notemos que "HS2\$HS" representa la información de la columna llamada "HS" y "HS2\$Sitios" representa la información de la columna llamada "Sitios". Las estadísticas descriptivas fueron aplicadas a los conjuntos de datos de cada sitio por separado. Con la función "tapply()" aplicamos los cálculos para cada conjunto de datos en arreglo por filas, mientras que las funciones "lapply()" y "sapply()" aplicamos los cálculos para datos en arreglo por columnas.

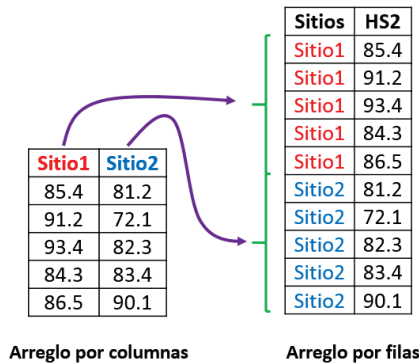


Figura 84. Diferencias en arreglo de los grupos de datos. A la izquierda los datos se encuentran arreglados por columnas; a la derecha los datos están arreglados por filas.

Regresando a la tabla de datos de “pH” y “humedad del suelo” llamada “Info_pH_HS”, a esta también podemos aplicar estadística descriptiva con funciones individuales para las dos columnas (pH y HS) utilizando “lapply()” y “sapply()”, a continuación calcularemos las medias (mean):

```
> lapply(Info_pH_HS[,2:3], FUN=mean)
$pH
[1] 5.08

$HS
[1] 82.04
> sapply(Info_pH_HS[,2:3], FUN=mean)
   pH    HS 
5.08 82.04
```

Las funciones “lapply()” y “sapply()” han calculado la media (promedio) de la columna 2 (pH) y 3 (HS) los cuales fueron 5.08 y 82.04 respectivamente. En este punto es notorio que la línea de código se ha tornado un tanto compleja, para recordar y estandarizar la nomenclatura de los objetos que conforman esa codificación, en la figura 85 se esquematizan las partes y los nombres de la línea de código utilizada arriba.

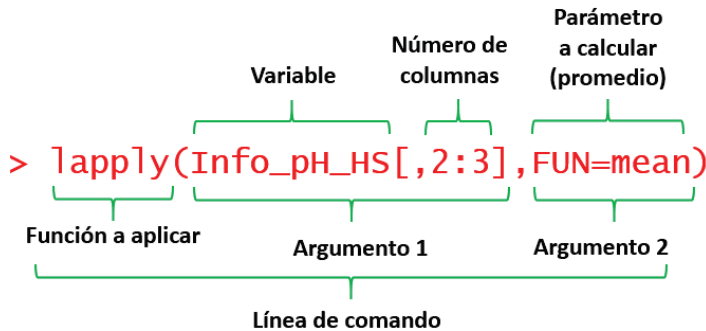


Figura 85. Línea de códigos utilizada para calcular el promedio de pH y HS que se encuentran en la tabla de datos guardada en la variable llamada "Info_pH_HS".

Así como calculamos la media, podemos calcular para ambas columnas las otras funciones de estadística descriptiva, por ejemplo la mediana, la desviación estándar, la cuenta, etc. (Cuadro 11).

La distribución normal

Necesitamos conocer si la distribución de los datos es normal, pues las pruebas paramétricas tienen como una premisa esta característica; si los datos no poseen dicha característica podemos seguir dos vías: 1. Transformar los datos o 2. Utilizar una prueba no paramétrica. Para que la distribución de los datos sea normal, estos deben seguir una distribución en forma de campana, como la observada en la figura 12, denominadas mesocúrtica y simétrica.

En R utilizaremos varias formas para explorar si los datos se distribuyen normalmente, algunas formas serán gráficas (histograma y gráfico Q), otras numéricas (curtosis y coeficiente de asimetría) y otras inferenciales (pruebas de Shapiro-Wilks y Kolmogorov-Smirnov). Adicionalmente se ofrecerá la opción de transformaciones para datos no normales.

Histograma

Un histograma es básicamente un gráfico de barra, en el cual se muestra en el eje X una variable numérica donde se distribuyen de forma ascendente los números o los rangos de números y en el eje Y se muestran las frecuencias de ocurrencias de los números o rangos de números.

Aplicaciones de Estadística Básica

En R podemos crear un histograma para un conjunto de datos con la función “hist()”. Para ejemplificarlo, haremos uso de una pequeña tabla de datos de la altura (en cm) medida a 10 personas. Primeramente estableceremos el directorio con la función “setwd()” y seguidamente los importaremos con el uso de la función “read.csv()” y guardaremos la información en una variable llamada “DatosAlt”:

```
> setwd("C:/Users/Miguel Garmendia/Documents/Archivos_R")
> dir()
[1] "DatosAlt.csv"
> DatosAlt <- read.csv("DatosAlt.csv")
> DatosAlt
```

	Personas	Altura
1	Persona1	143
2	Persona2	153
3	Persona3	160
4	Persona4	162
5	Persona5	165
6	Persona6	168
7	Persona7	172
8	Persona8	173
9	Persona9	177
10	Persona10	180

La función “dir()” es utilizada para visualizar los nombres de los archivos dentro del directorio (carpeta) que establecimos. Ahora los datos están listos para crear el histograma con la función “hist()”, solamente utilizando la columna denominada “Altura”:

```
> hist(DatosAlt$Altura)
```

En la figura 86 A se presenta el histograma resultante, el cual es totalmente básico y con nombres por defecto tanto en los ejes como en el título del gráfico, los cuales se pueden personalizar; sin embargo, antes de proceder con la personalización, cambiaremos el número de segmentaciones de datos en el eje X (número de rangos) utilizando el argumento “breaks=” dentro de la función “hist()” y estableceremos el número de segmentación en seis (Figura 86 B):

```
> hist(DatosAlt$Altura, breaks=6)
```

A como observamos, el número de barras no es exactamente seis, porque siempre habrá datos que no calzarán en la segmentación y el número de barras aumentará más de lo deseado, esto ocurre especialmente cuando tenemos pocos datos. Obviamente el lector

a su juicio y en dependencia de los datos que tenga, hará uso apropiado de esta función. A continuación haremos algunas personalizaciones a los gráficos utilizando varios argumentos diseñados para tales fines entre ellos “xlab=” para escribir el título del eje X; “ylab=” para escribir el título del eje Y; “main=” para escribir el título principal del gráfico, pero si no deseamos título principal escribimos “main=NULL” (Figura 86 C):

```
> hist(DatosAlt$Altura, xlab="Alturas", ylab="Frecuencias",  
main=NULL, breaks=6)
```

Observemos el cambio en los títulos de los ejes X y Y, y la ausencia del título principal del gráfico. Con una pieza más de código podemos cambiar el color de las barras, utilizando el argumento “col=” para hacerlo azul (blue) (Figura 86 D):

```
> hist(DatosAlt$Altura, xlab="Alturas", ylab="Frecuencias",  
main=NULL, breaks=6, col="blue")
```

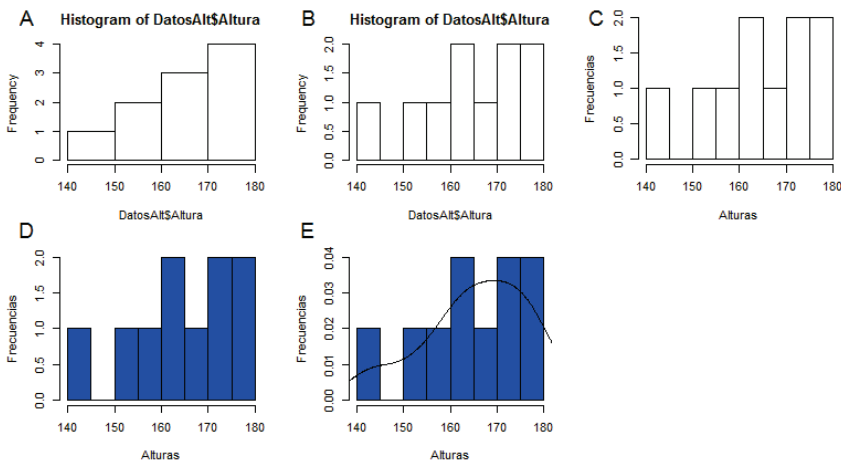


Figura 86. Histogramas representando cambios realizados mediante asignaciones de argumento a la función “hist()”. A. Apariencia del histograma por defecto; B. Aumento de la segmentación de los datos; C. Cambio en los nombres de los títulos de los ejes X y Y, y del título principal del gráfico; D. Asignación de color a las barras que conforman el histograma; E. Agregado de línea de densidad de las observaciones.

Adicionalmente podemos añadir una gráfica de línea que represente la densidad de las observaciones. Para ello, simplemente le indicamos al programa que no muestre en el gráfico la frecuencia, sino las probabilidades de densidad con el argumento “freq=FALSE”; seguidamente añadimos la línea con la función “lines()” (en una línea de comando aparte) y como argumento de la función, otra función llamada “density()” en la cual se asignan los datos a utilizar (Figura 86 E):

Aplicaciones de Estadística Básica

```
> hist(DatosAlt$Altura, xlab="Alturas", ylab="Frecuencias",  
main=NULL, breaks=6, col="blue", freq=FALSE)  
> lines(density(DatosAlt$Altura))
```

Es válido aclarar que no es objeto de este capítulo profundizar en las opciones de personalizaciones y de colores, ya que serán ampliamente abordadas posteriormente. Para los novicios en el uso de R, en la figura 87 se ofrece una explicación visual de los códigos del comando que originó la figura 86 D.

```
> hist(DatosAlt$Altura, xlab="Alturas", ylab="Frecuencias", main=NULL, breaks=6, col="blue")
```

<code>DatosAlt\$Altura,</code>	→	Indica la columna que se usará en el histograma.
<code>xlab="Alturas",</code>	→	Asigna el título "Alturas" al eje X.
<code>ylab="Frecuencias",</code>	→	Asigna el título "Frecuencias" al eje Y.
<code>main=NULL,</code>	→	Suprime el título principal al gráfico.
<code>breaks=6,</code>	→	Establece el número de segmentación del eje X.
<code>col="blue"</code>	→	Asigna color azul a las barras.

Figura 87. Ilustración de los argumentos dentro de la función "hist()" para personalizar el histograma.

El histograma permite observar cómo se distribuyen los datos, en especial si se distribuyen de forma normal o tienen otro tipo de distribución. En la figura 16 se presentan varias situaciones ilustradas donde se diferencia un histograma formado por datos provenientes de una distribución normal e histogramas con datos provenientes de otro tipo de distribución. Para el caso del ejemplo se sospecha que la distribución no es normal, pues las barras no forman una figura en forma de campana.

Gráfico Q

El gráfico Q o también llamado "Q-Q Plot" (Q-Q proviene del inglés Quantile Quantile), es una forma común de evaluar rápidamente si un grupo de datos se distribuyen de forma normal. R cuenta con una función directa para desplegar un gráfico Q, esta es la función "qqnorm()". Con la misma tabla de datos que utilizamos para demostrar el uso del histograma, ejemplificaremos el uso de la función "qqnorm()" (Figura 88 A):

```
> qqnorm(DatosAlt$Altura)
```

Sin embargo, si deseamos hacer algunas personalizaciones, podemos utilizar los mismos códigos de asignación de títulos usados en el histograma. En particular, la personalización que realizaremos a este gráfico serán: quitarle el título principal (Normal Q-Q Plot) y establecer el título "Observados" en el eje Y, "Teórico" en el eje X (Figura 88 B):

```
>qqnorm(DatosAlt$Altura, xlab="Teórico", ylab="Observados",  
main=NULL)
```

Adicionalmente con la función “qqline()” podemos agregar la línea típica de un gráfico Q (Figura 88 C).

```
>qqnorm(DatosAlt$Altura, xlab="Teórico", ylab="Observados",  
main=NULL)  
> qqline(DatosAlt$Altura)
```

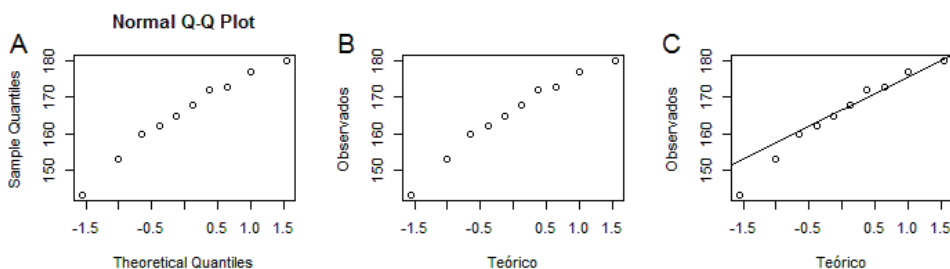


Figura 88. Gráfico Q representando cambios realizados mediante asignaciones de argumento a la función “qqnorm()”. A. Apariencia del gráfico Q por defecto; B. Cambio en los nombres de los títulos de los ejes X y Y, y del título principal del gráfico; C. Inclusión de la línea del gráfico Q.

Si los puntos se ubicaran muy cerca de la línea, entonces podríamos afirmar que los datos siguen una distribución normal. En la figura 22 se presentan varias situaciones de forma ilustrada en donde se diferencia un gráfico Q formado por datos provenientes de una distribución normal y datos provenientes de otro tipo de distribución. Para el caso del ejemplo, se sospecha que la distribución no es normal, pues algunos puntos se alejan considerablemente de la línea.

Método numérico (curtosis y coeficiente de asimetría)

Determinar la curtosis y el coeficiente de asimetría de un conjunto de datos es de vital importancia, pues nos da una idea sólida del tipo de distribución que siguen los datos. Dentro de sus funciones básicas, R no cuenta con una función directa para determinar la curtosis y el coeficiente de asimetría, de tal forma que tendríamos que instalar un paquete que contenga funciones para ejecutar dichos análisis, en este caso instalaremos el paquete llamado “e1071” mediante la función “install.packages()” o por medio de la opción de instalación explicada en la figura 81; seguidamente haremos disponible el paquete con la función “library()”:

Aplicaciones de Estadística Básica

```
> install.packages("e1071")  
> library("e1071")
```

Utilizaremos los mismos datos de altura de 10 personas que hemos estado utilizando en este acápite, junto con las funciones “kurtosis()” y “skewness()”, para calcular la curtosis y el coeficiente de asimetría respectivamente. A las funciones les debemos agregar dos argumentos adicionales además de los datos, el argumento “na.rm” el cual representa la pregunta: ¿Deberían ser removidos los valores perdidos?, la respuesta “FALSE”, indica que “no” y no responder o escribir “TRUE” responde a “sí”; y el argumento “type”, determina el tipo de algoritmo para realizar el cálculo, las opciones son:

Type 1= Algoritmo tradicional.

Type 2= Algoritmo utilizado en los programas SAS, SPSS y Microsoft Excel.

Type 3= Algoritmo utilizado por los programas Minitab y BMDP.

Para coincidir con los resultados del cálculo de la curtosis y el coeficiente de asimetría que realizamos en MS Excel, en este ejemplo utilizaremos el algoritmo tipo 2. Es conveniente aclarar que el lector es libre de revisar las otras opciones, las fórmulas de cada algoritmo aparecen en el manual del paquete, específicamente en Meyer et al. (2017). Las funciones “kurtosis()” para calcular la curtosis y “skewness()” para calcular el coeficiente de asimetría también se pueden encontrar en los paquetes “agricolae” y “moments”. La aplicación de las funciones y los argumentos en R se verían de la siguiente forma:

```
> kurtosis(DatosAlt$Altura, na.rm=FALSE, type=2)  
[1] 0.21724  
> skewness(DatosAlt$Altura, na.rm=FALSE, type=2)  
[1] -0.7268995
```

Si comparamos los datos, con los valores de referencia del cuadro 2, observamos que los datos forman una distribución más o menos “Platicúrtica” y evidentemente “Asimétrica negativa”. Con lo que concluimos que los datos aparentemente “no siguen una distribución normal”; sin embargo, serán las pruebas de inferencia las que tendrán la última palabra.

Método inferencial

Una serie de métodos inferenciales se han desarrollado para confirmar si un conjunto de datos siguen o no una distribución normal, los más comunes de encontrar en la literatura son, la prueba de Shapiro-Wilk con la función “shapiro.test()” y la prueba de Kolmogorov-Smirnov con la función “lillie.test()”, el primer método se suele aplicar si las observaciones son menos de 50 y el segundo si las observaciones son más de 50. El

lector está invitado a documentarse más sobre las pruebas y explorar sus diferencias con el uso de otros recursos, en esta publicación no se discutirán dichas diferencias. Otras pruebas para lograr los mismos objetivos son las de Anderson-Darling con la función “ad.test()”; Cramer-von Mises con la función “cvm.test()”; Chi-cuadrado de Pearson “pearson.test()” y Shapiro-Francia “sf.test()”. La hipótesis que siguen todas las pruebas son las siguientes:

H_0 : Las observaciones siguen una distribución normal.

H_1 : Las observaciones no siguen una distribución normal.

Como ejemplo utilizaremos los datos de altura de 10 personas usados anteriormente. En primer lugar se ejemplificará el uso de la prueba de Shapiro-Wilk:

```
> shapiro.test(DatosAlt$Altura)
```

```
Shapiro-Wilk normality test
```

```
data: DatosAlt$Altura  
W = 0.96004, p-value = 0.7863
```

Dado $\alpha = 0.05$, el valor de “p” (0.7863) es mayor que 0.05, por lo que no hay evidencias para afirmar que la hipótesis nula es falsa, de tal forma que concluimos que los datos siguen una distribución normal. A pesar de que con las otras opciones sospechábamos que los datos no se distribuían de forma normal, hemos comprobado con las pruebas de Shapiro-Wilk, que realmente sí se distribuyen de forma normal.

En caso que quisiéramos aplicar la prueba a varias columnas de datos, deberíamos hacer uso de la función “sapply()”. Por ejemplo, podemos aplicar la prueba a la columna de potencial de hidrógeno (pH) y a la de humedad del suelo (HS) del conjunto de datos guardados en la variable “Info_pH_HS”:

```
> Info_pH_HS  
PUNTO pH HS  
1 A 4.7 83.2  
2 B 5.3 95.4  
3 C 5.9 74.3  
4 D 4.9 87.6  
5 E 4.6 69.7  
  
> sapply(Info_pH_HS[,2:3], FUN=shapiro.test)  
pH HS  
statistic 0.9022805 0.9735313
```

Aplicaciones de Estadística Básica

```
p.value      0.4226178      0.8973968
method "Shapiro-Wilk normality test" "Shapiro-Wilk normality
test"
data.name "X[[i]]"      "X[[i]]"
```

Dado $\alpha = 0.05$, los valores de “p” para cada variable (0.4226178 y 0.8973968) son mayor que 0.05, por lo que no hay evidencias para afirmar que la hipótesis nula es falsa, de tal forma que concluimos que los datos de ambas variables siguen una distribución normal.

La segunda prueba a ejemplificar es la de Kolmogorov-Smirnov mediante la función “lillie.test()”; sin embargo, debemos instalar un paquete llamado “nortest”:

```
> install.packages("nortest")
> library("nortest")
```

Instalado el paquete, aplicamos la prueba a los datos de altura de las 10 personas:

```
> lillie.test(DatosAlt$Altura)

      Lilliefors (Kolmogorov-Smirnov) normality test

data:  DatosAlt$Altura
D = 0.12313, p-value = 0.9377
```

Dado $\alpha = 0.05$, el valor de “p” (0.9377) es mayor que 0.05, por lo que no hay evidencias para afirmar que la hipótesis nula es falsa, de tal forma concluimos que los datos siguen una distribución normal.

Transformaciones

Cuando determinamos que los datos no siguen una distribución normal, no es recomendado aplicar alguna prueba paramétrica cuya premisa (no la única) es que los datos deben seguir dicha distribución. Una manera de ajustar los datos que no siguen una distribución normal a una forma más o menos normal es transformarlos. Hay muchas formas de transformar los datos, entre ellas: Normal, referido a la normalización o estandarización de los datos, se realiza sustrayendo la media a cada dato y dividiendo el producto entre la desviación estándar muestral; Ln, o logaritmo neperiano; Log10, o logaritmo base 10; Log2, o logaritmo base 2; RaizCuad, que representa a la raíz cuadrada de los valores; Potencia, en el que se eleva a una potencia deseada; Inverso, se divide uno entre cada valor.

La transformación de los datos y su uso en R lo podemos hacer de dos formas, una es realizando las transformaciones en el programa MS Excel u otro programa estadístico y luego importando los resultados a R y otra es realizando las transformaciones directamente en R. Para transformar los datos en R utilizando logaritmo neperiano (ln), hacemos uso de la función “log()”, ejemplificaremos su uso con los datos de altura de personas que hemos venido utilizando:

```
> log(DatosAlt$Altura)
[1] 5.192957 5.147494 5.075174 5.153292 5.105945 5.123964
4.962845 5.087596
[9] 5.176150 5.030438
```

Si deseamos transformar y seguir manteniendo los datos transformados en un formato tabular, podemos hacer uso de la función “data.frame()” y guardar los datos en una variable a la que llamaremos “DatosAltLOG”:

```
> DatosAltLOG <-data.frame(log(DatosAlt$Altura))
> DatosAltLOG
  log.DatosAlt.Altura.
1          5.192957
2          5.147494
...
9          5.176150
10         5.030438
```

Ahora los datos están transformados y estructurado en un formato tabular, sin embargo necesitamos establecer un nombre más personalizado de la columna de los datos, esto lo logramos con el uso de la función “colnames()”, como argumento de la función escribimos la variable donde guardamos los datos y le asignamos el nombre (Alt_Log) entre comillas:

```
> colnames(DatosAltLOG) <- "Alt_Log"
> DatosAltLOG
  Alt_Log
1 5.192957
2 5.147494
...
9 5.176150
10 5.030438
```

Aplicaciones de Estadística Básica

En R podemos realizar otro tipo de transformaciones utilizando funciones definidas por el programa, por ejemplo: `log10()` que devuelve el valor del logaritmo base 10; `logb()` presenta el valor de un logaritmo de base personalizada, el argumento sería `logb(X,b)` donde X es el número a determinar el logaritmo y b es el número que representa la base; `exp()` eleva a la potencia; `sqrt()` devuelve la raíz cuadrada; entre otros.

Cuando precisamos realizar transformaciones a más de una columna de datos, los comandos varían un poco con respecto a los presentados anteriormente. Para ejemplificar esta variación, utilizaremos los datos de DAP (Diámetro a la Altura del Pecho) de seis árboles, utilizada anteriormente y le añadiremos una nueva variable correspondiente a la altura de cada árbol. Guardaremos la información en una variable llamada “DAP_Alt”:

```
> DAP_Alt <-read.csv(file.choose())
```

```
> DAP_Alt
```

	Arbol	Sitio	DAP	Altura
1	1	BS	37.2	16.4
2	2	BS	23.1	7.9
3	3	BS	12.3	6.5
4	4	BP	59.6	19.6
5	5	BP	89.6	45.5
6	6	BP	65.9	24.6

A continuación, se mostrarán dos formas de ejecutar la transformación de los datos de las dos columnas (DAP y Altura). En la primera crearemos dos vectores de los datos transformados por separados y los uniremos en una tabla de datos, entonces aplicamos la función “`log()`” a la columna llamada DAP y su resultado lo grabamos en la variable “DAP_Log” y seguidamente aplicamos la misma función a la columna Altura y grabamos su resultado en la variable “Alt_Log”, posteriormente unimos las dos variables con el uso de la función “`data.frame()`” y el resultado global lo guardamos en una nueva variable, a la que denominaremos “DAP_Alt2”. Primeramente deberemos crear dos variables con la información del logaritmo calculado y luego fusionamos las dos columnas, con el uso de la función “`data.frame()`”:

```
> DAP_Log <-log(DAP_Alt$DAP)
```

```
> Alt_Log <-log(DAP_Alt$Altura)
```

```
> DAP_Alt2 <-data.frame(DAP_Log, Alt_Log)
```

```
> DAP_Alt2
```

	DAP_Log	Alt_Log
1	3.616309	2.797281
2	3.139833	2.066863
3	2.509599	1.871802

En Microsoft® Excel y R

```
4 4.087656 2.975530
5 4.495355 3.817712
6 4.188138 3.202746
```

Notemos que el nombre que se le asigna a la columna en la nueva tabla de datos (DAP_Log y Alt_Log) son los mismos nombres de las variables en las que se guardaron los datos transformados respectivamente.

La segunda forma es más sencilla, pues utilizamos la función “log()” directamente con la selección de las columnas (la variable DAP está en la columna 3 y la variable Altura está en la columna 4) de datos a utilizar en la transformación:

```
> log(DAP_Alt[,3:4])
      DAP      Altura
1 3.616309 2.797281
2 3.139833 2.066863
3 2.509599 1.871802
4 4.087656 2.975530
5 4.495355 3.817712
6 4.188138 3.202746
```

Podemos guardar los resultados de la transformación en alguna variable para posterior uso. En R existen muchas formas de poder hacer una misma tarea, estas han sido solamente dos formas y el lector es libre de utilizar la que más le interese.

Estadística inferencial

A como se explicó en MS Excel, la idea con la estadística inferencial es generar un valor probabilístico para “rechazar” o “no rechazar” (y por consiguiente aceptar) una hipótesis. Este capítulo está dividido en tres temas, la comparación de proporciones y frecuencias, la comparación de medias y la búsqueda de relaciones.

Comparación de proporciones y frecuencias

Estas comparaciones son útiles, en especial cuando se trabaja con información cualitativa. Una proporción es un valor relativo de un dato en relación al total del conjunto de datos al que pertenece. Las proporciones se pueden expresar en porcentaje al multiplicarse por 100 y la sumatoria de ellas debe ser igual a 1. Por otro lado, la frecuencia es el número de veces que se cuenta un objeto, evento, organismo o número. Ejemplo: Número de huevos en un nido; número de plantas dañadas por un hongo; número de personas de tamaño entre 1.5 y 1.8 metros, número de veces que una persona responde “Sí”

Aplicaciones de Estadística Básica

o “NO” en una encuesta, etc. Es común que las frecuencias se expresen en proporciones. Esta sección está dividida en: Prueba de una proporción; prueba de dos proporciones; bondad de ajuste; y pruebas de independencia, aplicadas explícitamente para análisis de frecuencia donde se prueba la dependencia o independencia entre 2 x 2 condiciones, o entre R x C condiciones.

Prueba de una proporción

La prueba de una proporción es útil para comparar una proporción obtenida mediante muestreo o de forma experimental (observada) y otra proporción preestablecida (teórica). La prueba de una proporción asume las siguientes hipótesis:

$$H_0: \hat{p} = p$$

$$H_1: \hat{p} \neq p; \text{ también llamada de dos colas.}$$

$$H_1: \hat{p} > p; \text{ también llamada de una cola hacia la derecha.}$$

$$H_1: \hat{p} < p; \text{ también llamada de una cola hacia la izquierda.}$$

Donde,

\hat{p} = Proporción observada.

p = Proporción teórica.

La función “prop.test()” se utiliza ampliamente en R para hacer los cálculos de proporciones. Los argumentos deben incluir el número de observaciones acertadas (x), el número total de observaciones (n) y la proporción teórica (pt), de tal forma que el comando completo quedaría estructurado como: “prop.test(x,n,pt)”. En dependencia del tipo de prueba de proporciones (una o dos colas), así se utiliza la función en R (Cuadro 13).

Cuadro 13. Tipo colas y funciones + argumentos asociadas para lograr los cálculos.

COMPARACIONES	FUNCIÓN + ARGUMENTOS
Una cola - hacia la derecha	prop.test(x, n, pt, alternative=“greater”)
Una cola - hacia la izquierda	prop.test(x, n, pt, alternative=“less”)
Dos colas	prop.test(x, n, pt)

x= observaciones acertadas; n= total de observaciones; pt= proporción teórica. El argumento “alternative” (“alternativo” en español) define el tipo de cola.

Para ejemplificar la prueba de una proporción con una cola hacia la derecha, utilizaremos la siguiente situación hipotética: En un estudio de sanidad de peces en un río se presume que el número de peces afectados por un parásito es mayor al 10%. Para probar si este valor es correcto, se realiza un muestreo aleatorio de peces donde se revisan

756 peces, de los cuales 134 estaban afectados por el parásito. Se pretende determinar si la proporción de los datos muestreados, realmente coincide con la hipótesis $>10\%$ asumida:

```
> prop.test(134, 756, 0.1, alternative="greater")

1-sample proportions test with continuity correction

data: 134 out of 756, null probability 0.1
X-squared = 49.271, df = 1, p-value = 1.115e-12
alternative hypothesis: true p is greater than 0.1
95 percent confidence interval:
 0.1549406 1.0000000
sample estimates:
               p
0.1772487
```

El valor de “p”(p-value) es igual a 1.115×10^{-12} , lo que es equivalente a decir 1.115×10^{-12} ; dado $\alpha = 0.05$, el valor de “p” es extremadamente menor a 0.05, por lo que tenemos evidencias para rechazar la hipótesis nula y concluir que realmente el número de peces afectados por un parásito es mayor al 10%. Notemos que el estadístico se representa como “X-squared” y en este caso es igual a 49.271; además, calcula un intervalo de confianza del 95%, finalmente presenta el valor de la proporción. El lector es libre de sacar provecho a la información resultante de la aplicación de la prueba con la función “prop.test()”, y está invitado a utilizar otros recursos para aclarar dudas concernientes a dicha información.

Para ejemplificar la prueba de una proporción con una cola hacia la izquierda, se utilizará la siguiente situación hipotética: En un estudio de mortalidad de plantas de una especie de árbol sembrada en un plan de reforestación, se presume que la mortalidad de las mismas es menor del 20%. Pasado un lapso de tiempo después de haberlas sembrado, y para confirmar que la mortalidad realmente era menor al 20%, se realizó un muestreo de 1393 plantas y se verificó que 102 estaban muertas. De tal forma que es interés del estudio el comparar la proporción de plantas muestreadas con la proporción teórica:

```
> prop.test(102, 1393, 0.2, alternative="less")

1-sample proportions test with continuity correction

data: 102 out of 1393, null probability 0.2
X-squared = 139.14, df = 1, p-value < 2.2e-16
```

Aplicaciones de Estadística Básica

```
alternative hypothesis: true p is less than 0.2
95 percent confidence interval:
 0.00000000 0.08593378
sample estimates:
      p
0.07322326
```

El valor de “p” (p-value) es igual a 2.2×10^{-16} , lo que es equivalente a decir 2.2×10^{-16} ; dado $\alpha = 0.05$, el valor de “p” es extremadamente menor a 0.05, por lo que se tiene evidencias para rechazar la hipótesis nula y se concluye que realmente el número de plantas muertas es mejor de 20%.

La prueba de una proporción para dos colas será ejemplificada con la siguiente situación hipotética: Se realiza una encuesta para determinar qué porcentaje de productores en un municipio están aplicando obras de conservación de suelo, estudios anteriores afirman que el 50% de ellos lo están haciendo. De 46 encuestas aplicadas, se determina que 22 respondieron positivamente a la opción que refleja la aplicación de obras de conservación. Se pretende comparar la proporción muestreada con la descrita por los estudios anteriores.

```
> prop.test(22, 46, 0.5)

1-sample proportions test with continuity correction

data:  22 out of 46, null probability 0.5
X-squared = 0.021739, df = 1, p-value = 0.8828
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3313667 0.6287485
sample estimates:
      p
0.4782609
```

Dado $\alpha = 0.05$, el valor de “p” (0.8828) es mayor que 0.05, por lo que se falla de rechazar la hipótesis nula y se concluye que realmente la proporción de encuestados que están haciendo obras de conservación de suelo, es igual al reportado por otros estudios (50%).

Pruebas de dos proporciones

Estas pruebas son útiles para comparar dos proporciones que generalmente son provenientes de experimentos o muestreos (observaciones), asumiendo las siguientes hipótesis:

$$H_0: \hat{p}_1 = \hat{p}_2$$

$$H_1: \hat{p}_1 \neq \hat{p}_2; \text{ también llamada de dos colas.}$$

$$H_1: \hat{p}_1 > \hat{p}_2; \text{ también llamada de una cola hacia la derecha.}$$

$$H_1: \hat{p}_1 < \hat{p}_2; \text{ también llamada de una cola hacia la izquierda.}$$

Donde,

\hat{p}_1 = Proporción uno.

\hat{p}_2 = Proporción dos.

De nuevo se hará uso de la función “prop.test()” para hacer los cálculos de proporciones en R. Los argumentos deben incluir para la primera proporción: el número de observaciones acertadas (x1) y el número total de observaciones (n1), y para la segunda proporción: el número de observaciones acertadas (x2), el número total de observaciones (n2), de tal forma que el comando completo quedaría estructurado como: “prop.test(c(x1,x2), c(n1,n2))”. En dependencia del tipo de prueba de proporciones (una o dos colas), así se utiliza la función en R (Cuadro 14).

Cuadro 14. Tipo colas y funciones + argumentos asociadas para lograr los cálculos.
comparaciones Función + argumentos

COMPARACIONES	FUNCIÓN + ARGUMENTOS
Una cola - hacia la derecha	prop.test(c(x1,x2), c(n1,n2), alternative=“greater”)
Una cola - hacia la izquierda	prop.test(c(x1,x2), c(n1,n2), alternative=“less”)
Dos colas	prop.test(c(x1,x2), c(n1,n2))

x1= observaciones acertadas de la muestra 1; x2= observaciones acertadas de la muestra 2; n1= total de observaciones en la muestra 1; n2= total de observaciones en la muestra 2. El argumento “alternative” (“alternativo” en español) define el tipo de cola.

Para ejemplificar la prueba de dos proporciones de dos colas, se utilizará la siguiente situación hipotética: Se realizan encuestas en dos comunidades rurales para determinar si las proporciones de aceptación de una reforma a la ley ambiental, es similar entre ellas. En la comunidad 1 se aplicaron 77 encuestas, de las cuales 34 encuestados afirman están de acuerdo; en la comunidad 2 se aplicaron 68 encuestas, de las cuales 56 encuestados están de acuerdo. Se pretende comparar las proporciones de encuestados de ambas comunidades que están de acuerdo con la propuesta.

Aplicaciones de Estadística Básica

```
> prop.test(c(34,56), c(77,68))
```

```
2-sample test for equality of proportions with continuity
correction
```

```
data: c(34, 56) out of c(77, 68)
X-squared = 20.785, df = 1, p-value = 5.138e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5390368 -0.2249051
sample estimates:
   prop 1    prop 2 
0.4415584 0.8235294
```

Dado $\alpha = 0.05$, el valor de “p” ($5.138e-06$) es mucho menor que 0.05, por lo que se rechaza la hipótesis nula y se concluye que hay diferencias significativas entre la proporción de personas que dijeron estar de acuerdo con la reforma a la ley ambiental en la comunidad 1 con respecto a la proporción determinada en la comunidad 2. Y comparando las proporciones (prop 1 = 0.44, prop 2 = 0.82) se concluye que en la comunidad 2 la aceptación fue mayor.

Prueba de bondad de ajuste

Esta prueba se utiliza para determinar si un conjunto de frecuencias observadas se ajustan a un conjunto de proporciones predefinidas. La prueba de bondad de ajuste, asume las siguientes hipótesis:

H_0 : La variable aleatoria sigue la distribución conocida.

H_1 : La variable aleatoria sigue una distribución diferente.

O sea, la hipótesis nula (H_0) asume que los datos observados se ajustan a la distribución sugerida; y la hipótesis alternativa (H_1) supone que los datos observados no se ajustan a la distribución sugerida y tienen diferente distribución.

La función que R posee para realizar esta prueba es “chisq.test()” y como argumento se deben especificar las frecuencias observadas (Obser) y las proporciones (Prop), de tal forma que el comando quedaría organizado como: chisq.test(Obser, p= Prop).

Para ejemplificar la prueba de bondad de ajuste utilizaremos la siguiente situación hipotética: Se conoce que las semillas de una planta que produce frutos para la alimentación humana al ser sembradas producirán cuatro condiciones de frutos: el 56% de las

plantas que crezcan producirán frutos grandes y sin semillas, el 19% tendrán frutos grandes y con semillas, el 19% producirá frutos pequeños y con semillas y solamente un 6% producirá frutos pequeños y sin semillas.

El propietario de una finca de grandes dimensiones donde se cultivan dichas plantas quiere determinar si las plantas están produciendo frutos con las características y proporciones que le había ofrecido el vendedor de las semillas. Entonces, se realiza un muestreo y de forma aleatoria se seleccionan 5667 plantas, se miden los frutos y se revisa la presencia o ausencia de semillas, de tal forma que de las 5667 plantas: 3200 produjeron frutos grandes y sin semillas, 1057 frutos grandes y con semillas, 1072 frutos pequeños y con semillas, y 338 frutos pequeños y sin semillas. Se precisa saber si las frecuencias encontradas en el muestreo se ajustan a las proporciones dadas teóricamente.

Los datos se han tabulados en una hoja de cálculo y se importarán a R, las proporciones representadas por porcentajes se han dividido entre 100 previamente, los datos se verían en R de la siguiente forma:

```
> Semillas <-read.csv(file.choose())
> Semillas
  Obser Prop
1  3200 0.56
2  1057 0.19
3  1072 0.19
4   338 0.06
```

Como argumentos de la función “chisq.test()” se definen los datos observados y las proporciones esperadas:

```
> chisq.test(Semillas$Obser, p=Semillas$Prop)

Chi-squared test for given probabilities

data:  Semillas$Obser
X-squared = 0.61526, df = 3, p-value = 0.8929
```

Dado $\alpha = 0.05$, el valor de “p” es mucho mayor que 0.05, por lo que se falla en rechazar la hipótesis nula y se concluye que las frecuencias observadas se ajustan a las proporciones teóricas que denotaban las características de los frutos en las plantas en cuestión.

Aplicaciones de Estadística Básica

Pruebas de independencia (tablas de contingencia)

Estas pruebas determinan la relación entre variables utilizando datos de frecuencias. En este escrito, las pruebas se dividirán en dos, las de 2×2 y las de $R \times C$. Ambos tipos de tablas utilizan la prueba Chi-Cuadrado (χ^2) y similar procedimiento.

Tablas de contingencia 2×2

La prueba compara dos factores con dos niveles, o dos variables con dos niveles cada una, asumiendo las siguientes hipótesis:

H_0 : Las dos variables son independientes.

H_1 : Las dos variables son dependientes.

O sea, la hipótesis nula (H_0) asume que los dos factores o variables no tienen ninguna relación; y la hipótesis alternativa (H_1) supone que los dos factores o variables si tienen relación.

Para ejemplificar la prueba de independencia, utilizaremos la siguiente situación hipotética: Se realizan encuestas a dos comunidades sobre el “estar o no estar de acuerdo” con el establecimiento de una empresa en sus territorios. Sin embargo, existe la sospecha de que, por alguna razón intrínseca, el responder “SÍ” o “NO” depende de las comunidades. Para ello se aplicaron 77 encuestas en la comunidad 1 y 68 en la comunidad 2. En la comunidad 1, 34 personas respondieron SÍ y 43 respondieron NO; en la comunidad 2, 56 personas respondieron SÍ y 12 seleccionaron NO. Se pretende determinar si hay una relación entre las respuestas y las comunidades.

El cálculo del estadístico y el valor de “p” lo realizaremos con la función “chisq.test()” y los datos los arreglaremos en formato vectorial, en una variable guardaremos las frecuencias de las respuestas “SÍ” y en otra las frecuencias de las respuestas “NO”; adicionalmente, haremos uso de la función data.frame() para transformar los vectores en formato tabular al aplicar la función:

```
> Si <-c(34,56)
> No <-c(43,12)
> chisq.test(data.frame(Si, No))
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: data.frame(Si, No)
X-squared = 20.785, df = 1, p-value = 5.138e-06
```

Dado $\alpha = 0.05$, el valor de “p” (5.138e-06) es mucho menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que las variables son dependientes (tienen relación), por lo tanto el responder “SÍ” o “NO” está significativamente en dependencia del tipo de comunidad (1 o 2).

Tablas de contingencia R x C

La prueba compara más de dos factores, con más de dos niveles; o dos variables con más de dos niveles cada una, asumiendo las siguientes hipótesis:

H_0 : Las variables son independientes.

H_1 : Las variables son dependientes.

O sea, la hipótesis nula (H_0) asume que los factores o variables no tienen ninguna relación; y la hipótesis alternativa (H_1) supone que los factores o variables si tienen relación.

Para ejemplificar la prueba de independencia, utilizaremos la siguiente situación hipotética: Se realizan encuestas a tres comunidades sobre el “estar o no estar de acuerdo” con el establecimiento de una empresa en sus territorios. Sin embargo, existe la sospecha de que, por alguna razón intrínseca, el responder “SÍ”, “NO” o “Indiferente” (o sea que no le importa) depende de las comunidades. Para ello se aplicaron 83 encuestas en la comunidad 1, 76 en la comunidad 2 y 60 en la comunidad 3. En la comunidad 1, 34 respondieron “SÍ”, 43 “NO” y 6 “Indiferente”; en la comunidad 2, 56 respondieron “SÍ”, 12 “NO” y 8 “Indiferente”; y en la comunidad 3, 12 respondieron “SÍ”, 38 “NO” y 10 “Indiferente”. Se pretende determinar si hay una relación entre las respuestas y las comunidades.

El cálculo del estadístico y el valor de “p” lo realizaremos con la función “chisq.test()” y los datos los arreglaremos en formato vectorial, en una variable guardaremos las frecuencias de las respuestas “SÍ”, en otra las frecuencias de las respuestas “NO” y en otra las de la respuesta “Indiferente”; adicionalmente, haremos uso de la función “data.frame()” para transformar los vectores en formato tabular al aplicar la función:

```
> Si <-c(34,56,12)
> No <-c(43,12,38)
> Indiferente <-c(6,8,10)
> chisq.test(data.frame(Si,No,Indiferente))
```

Pearson's Chi-squared test

```
data: data.frame(Si, No, Indiferente)
X-squared = 45.095, df = 4, p-value = 3.799e-09
```

Aplicaciones de Estadística Básica

Dado $\alpha = 0.05$, el valor de “p” (3.799e-09) es mucho menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que las variables son dependientes (tienen relación), por lo tanto el responder “SÍ”, “NO” o “Indiferente” está significativamente en dependencia del tipo de comunidad (1, 2 o 3).

Comparación de medias

A como su nombre lo indica, el objetivo de este conjunto de pruebas es el comparar las medias (promedios) de conjuntos de valores numéricos en la mayoría de los casos continuos. En este escrito, se describirá la aplicación de dos tipos de pruebas, una prueba para comparar las medias de dos grupos y una prueba para comparar las medias de más de dos grupos. En específico se abordará la “prueba T” y el “análisis de varianza”. Para ejecutar la prueba T y el análisis de varianza es necesario confirmar el supuesto igualdad de varianza, por lo que en esta sección también se incluirá la prueba F.

Confirmar la igualdad de varianza es imprescindible para ejecutar distintos tipos de comparaciones de medias, afortunadamente R tiene una función que hace esta comprobación de una forma rápida y sencilla. La prueba F asume las hipótesis:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Donde,

σ_1^2 = Varianza del grupo 1.

σ_2^2 = Varianza del grupo 2.

Para ejemplificar el uso de la prueba F, rescataremos el ejemplo que utilizamos para el mismo tipo de prueba en MS Excel, en un estudio de suelo se ha medido la humedad (%) en 10 puntos aleatorios en dos sitios. Entonces, como resultado tenemos dos grupos de datos, los datos del sitio 1 y los datos del sitio 2 de una sola variable (humedad del suelo en %). Esta información primeramente la importamos a R y la guardaremos en la variable “HS”:

```
> HS <- read.csv(file.choose())
> HS
```

	No.Observ	Sitio1	Sitio2
1	1	85.4	81.2
2	2	91.2	72.1
...			
9	9	96.1	81.2
10	10	83.2	82.3

Para ejecutar la prueba, utilizaremos la función “var.test()” que generará un valor de probabilidad (p) con la cual se podrá rechazar o fallar en rechazar H_0 :

```
> var.test(HS$Sitio1, HS$Sitio2)

F test to compare two variances

data:  HS$Sitio1 and HS$Sitio2
F = 1.4196, num df = 9, denom df = 9, p-value = 0.6101
alternative hypothesis: true ratio of variances is not equal
to 1
95 percent confidence interval:
 0.3526041 5.7152289
sample estimates:
ratio of variances
      1.419582
```

Dado $\alpha = 0.05$, el valor de “p” (0.6101) es mayor que 0.05, por lo que no hay suficientes evidencias para rechazar H_0 y concluimos que las varianzas son iguales. Podemos asignar el tipo de cola en la función añadiendo el argumento “alternative=” y seleccionando el tipo de cola, dos colas: “two.sided”, una cola hacia la izquierda: “less”, una cola hacia la derecha: “greater”. Si no especificamos las colas, el programa hará los cálculos con la opción de dos colas por defecto. Para el ejemplo, si el comando quedara escrito como: `var.test(HS$Sitio1,HS$Sitio2, alternative=“greater”)` el resultado que producirá es el mismo producido por la opción homóloga en MS Excel.

Prueba T para una muestra

Esta prueba compara la media de un conjunto de valores con un valor teórico preestablecido. La prueba T para una muestra asume las siguientes hipótesis:

$$H_0: \bar{x} = \mu$$

$H_1: \bar{x} \neq \mu$; también llamada de dos colas.

$H_1: \bar{x} > \mu$; también llamada de una cola hacia la derecha.

$H_1: \bar{x} < \mu$; también llamada de una cola hacia la izquierda.

Donde,

\bar{x} = Media del conjunto de datos.

μ = Dato teórico.

Aplicaciones de Estadística Básica

Para aplicar la prueba, se hará uso de la función “t.test()”. Los argumentos serían “x” que representa el conjunto de datos y “y” que representa el valor teórico; de tal forma que el comando para el análisis quedaría expresado como: `t.test(x, mu= y)`, donde “mu” es igual a “y”. En dependencia del tipo de prueba T (una o dos colas), así se utiliza la función en R (Cuadro 15).

Cuadro 15. Tipo de colas y las fórmulas + funciones asociadas para lograr los cálculos.

COMPARACIONES	FUNCIÓN + ARGUMENTOS
Una cola – hacia la izquierda	<code>t.test(x, mu=y, alternative="less")</code>
Una cola – hacia la derecha	<code>t.test(x, mu=y, alternative="greater")</code>
Dos colas	<code>t.test(x, mu=y)</code>

x= el conjunto de datos (grupo); mu= y; y= el valor teórico. El argumento “alternative” (“alternativo” en español) define el tipo de cola.

Para ejemplificar la prueba con cola hacia la izquierda, se utilizará la siguiente situación hipotética: Los niveles de oxígeno disuelto (OD) en el agua no pueden ser menores de 3 ppm (partes por millón), de serlo toda la fauna acuática estaría en peligro. A lo largo de una fuente de agua se tomaron siete muestras y se determinó el OD a cada una, de tal forma que se pretende comparar la media de los datos observados con el valor de referencia (<3 ppm), para determinar si los niveles de OD realmente son menores a ese valor o no. Primeramente importamos los datos a R:

```
> OD <- read.csv(file.choose())
> OD
  Muestras  OD
1         1  2.2
2         2  3.2
3         3  2.1
4         4  2.3
5         5  3.1
6         6  2.5
7         7  1.3
```

Seguidamente aplicamos la prueba de una cola hacia la izquierda a los datos de OD:

```
> t.test(OD$OD, mu=3, alternative="less")
```

One Sample t-test

```
data: OD$OD
t = -2.5236, df = 6, p-value = 0.04507
alternative hypothesis: true mean is not equal to 3
```

```
95 percent confidence interval:  
 1.790095 2.981334  
sample estimates:  
mean of x  
 2.385714
```

Dado $\alpha = 0.05$, el valor de “p” es muy cercano a 0.05, lo que nos sugiere que los valores están en el borde del rechazo de la hipótesis nula. Dado a que no podemos llegar a una conclusión definitiva, es el investigador quién decide si “estar al borde del rechazo” traerá consigo consecuencias biológicas a los organismos o no.

El abordaje para las pruebas de dos colas y de una cola hacia la derecha es similar al descrito en el ejemplo anterior y las diferencias radican en el uso de la opción “alternativa” del comando, dichas variaciones se muestran en el cuadro 15.

Prueba T para dos muestras independientes

Esta prueba compara dos conjuntos (grupos) de datos de la misma variable y genera un valor de significancia que nos sirve para decidir si los conjuntos de datos son semejantes o diferentes. La prueba T para dos muestras independientes asume las siguientes hipótesis:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$; también llamada de dos colas.

$H_1: \mu_1 > \mu_2$; también llamada de una cola hacia la derecha.

$H_1: \mu_1 < \mu_2$; también llamada de una cola hacia la izquierda.

Donde,

μ_1 = Media del conjunto de datos 1.

μ_2 = Media del conjunto de datos 2.

Retomaremos la función “t.test()” para aplicar la prueba. Dentro de la función se especificarán los argumentos “x1” que representa el conjunto de datos 1 y “x2” que representa el conjunto de datos 2; de tal forma que el comando para el análisis quedaría expresado como: t.test(x1, x2). En dependencia del tipo de prueba T (una o dos colas), así se utiliza la función en R (Cuadro 16).

Cuadro 16. Tipo colas y las fórmulas y funciones asociadas para lograr los cálculos.

COMPARACIONES	FUNCIÓN + ARGUMENTOS
Una cola – hacia la izquierda	<code>t.test(x1, x2, alternative="less")</code>
Una cola – hacia la derecha	<code>t.test(x1, x2, alternative="greater")</code>
Dos colas	<code>t.test(x1, x2)</code>

x_1 = el conjunto de datos (grupo) 1; x_2 = el conjunto de datos 2. El argumento "alternative" ("alternativo" en español) define el tipo de cola.

Para ejemplificar el uso de la prueba T retomaremos los datos de humedad del suelo utilizados anteriormente. La pregunta sería: ¿Qué tan diferentes son los dos grupos de datos? Si asumimos que la igualdad de varianzas es verdadera, se aplicará la función de la siguiente manera:

```
> t.test(HS$Sitio1, HS$Sitio2, var.equal=TRUE)
```

```
Two Sample t-test
```

```
data: HS$Sitio1 and HS$Sitio2
t = 3.2094, df = 18, p-value = 0.004861
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 3.056646 14.643354
sample estimates:
mean of x mean of y
 88.98    80.13
```

Notemos que el argumento "var.equal=TRUE" lo incluimos, cuando se asume igualdad de varianza; si no asumimos igualdad de varianza, entonces no expresamos el argumento "var.equal=" o lo expresamos como "var.equal=FALSE".

Dado $\alpha = 0.05$, el valor de "p" es menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que la humedad del suelo es significativamente diferente entre los dos sitios. El abordaje para las pruebas de una cola hacia la derecha e izquierda es similar al descrito en el ejemplo anterior y las diferencias radican en el uso de la opción "alternative" del comando, dichas variaciones se muestran en el cuadro 16.

Prueba T no paramétrica para dos muestras independientes (Wilcoxon o Mann-Whitney)

Se utiliza para comparar las medias de dos conjuntos de datos, para los cuales se ha confirmado que al menos uno de ellos no se distribuye normalmente. De tal forma que esta es una prueba no paramétrica equivalente a la prueba T paramétrica. La hipótesis que se asume es la misma para la prueba paramétrica. El tipo y dirección de las colas también se determinan a como se han determinado anteriormente en la prueba T paramétrica. Para aplicar esta prueba no paramétrica utilizaremos la función “wilcox.test()”. Se utilizarán, a modo de ejemplificación, los mismos datos de humedad de suelo utilizados anteriormente, dicha información será importada en R, guardada en la variable “HS” y se le aplicará la prueba:

```
> HS <-read.csv(file.choose())  
> wilcox.test(HS$Sitio1, HS$Sitio2)
```

```
Wilcoxon rank sum test with continuity correction
```

```
data: HS$Sitio1 and HS$Sitio2  
W = 88, p-value = 0.004556  
alternative hypothesis: true location shift is not equal to 0
```

Dado $\alpha = 0.05$, el valor de “p” es menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que la humedad relativa entre los dos sitios es significativamente diferente.

Prueba T para dos muestras pareadas

La prueba T para dos muestras pareada compara dos conjuntos de datos (grupos) que son dependientes entre ellos, pues contienen medidas que se han tomado a los mismos objetos en el tiempo (generalmente). La prueba T para dos muestras pareadas, asume las siguientes hipótesis:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0; \text{ también llamada de dos colas.}$$

$$H_1: \mu_d > 0; \text{ también llamada de una cola hacia la derecha.}$$

$$H_1: \mu_d < 0; \text{ también llamada de una cola hacia la izquierda.}$$

Donde,

μ_d = La media de la diferencia.

d = Diferencia entre la primera medida y la segunda medida.

Aplicaciones de Estadística Básica

Se hará uso de la función “t.test()” para aplicar la prueba. Dentro de la función especificaremos los argumentos “m1” que representa el conjunto de datos del momento 1 y “m2” que representa el conjunto de datos del momento 2; además se asignaremos el argumento “paired=TRUE” para indicarle al programa que la prueba T a ejecutar es pareada, de tal forma que el comando para el análisis se escribiría de la siguiente manera: t.test(m1, m2, paired=TRUE). En dependencia del tipo de prueba T (una o dos colas), así se utiliza la función en R (Cuadro 17).

Cuadro 17. Tipo colas y las fórmulas y funciones asociadas para lograr los cálculos.

COMPARACIONES	FUNCIÓN + ARGUMENTOS
Una cola – hacia la izquierda	t.test(m1, m2, paired=TRUE, alternative=“less”)
Una cola – hacia la derecha	t.test(m1, m2, paired=TRUE, alternative=“greater”)
Dos colas	t.test(m1, m2, paired=TRUE)

m1= el conjunto de datos del momento 1; m2= el conjunto de datos del momento 2. El argumento “paired=TRUE” le indica que la prueba es pareada. El argumento “alternative” define el tipo de cola.

Para ejemplificar, utilizaremos una situación hipotética donde asumiremos que se ha medido el peso (lb) de un grupo de 10 venados antes y después, de haber suministrado dosis continuas de desparasitante para parásitos internos en un plan de Manejo de Fauna Silvestre. Los venados estaban codificados y uno a uno se capturó y se midió su peso inicial. Tres meses después de estar suministrando el desparasitante, en los bebederos artificiales, se volvieron a capturar y a pesar los mismos 10 venados. El objetivo es conocer si el peso de los venados “antes y después” cambió significativamente, dicho cambio podría ser atribuido (posiblemente) al suministro del desparasitante. Primero importamos los datos y los guardamos en una variable llamada “Venados”:

```
> Venados <-read.csv(file.choose())
```

```
> Venados
```

```
  Antes Despues
1   78.5     84.2
2  123.2    130.4
3  150.8    151.3
4  121.3    122.3
5   79.8     86.7
6   98.5    110.5
7   89.3     89.4
8  102.5    109.4
9  145.6    151.9
10  89.9     97.7
```

Seguidamente aplicamos la prueba estadística para datos pareados:

```
> t.test(Venados$Antes, Venados$Despues, paired=TRUE)
```

Paired t-test

```
data: Venados$Antes and Venados$Despues
t = -4.5368, df = 9, p-value = 0.001412
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -8.152535 -2.727465
sample estimates:
mean of the differences
      -5.44
```

Dado $\alpha = 0.05$, el valor de “p” es menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que si hay diferencia entre los pesos de los venados de un momento al otro. El abordaje para las pruebas de una cola hacia la derecha o izquierda es similar al descrito en el ejemplo anterior y las diferencias radican en el uso de la opción “alternative” del comando, dichas variaciones se muestran en el cuadro 17.

Prueba T no paramétrica para dos muestras pareadas (Wilcoxon)

Se utiliza para comparar las medias de dos conjuntos de datos pareados, para los cuales se ha confirmado que al menos uno de ellos no se distribuye normalmente. Las hipótesis que se asumen, el tipo y dirección de las colas se determinan a como se han determinado anteriormente en las pruebas T paramétricas. Utilizaremos la función “wilcox.test()” con el argumento “paired=TRUE” para aplicar esta prueba. Haremos uso de los datos de los venados para ejemplificar su aplicación, entonces importamos la información a R y la guardamos en una variable llamada “Venados”, consecutivamente aplicamos la prueba utilizando el argumento “paired=TRUE”:

```
> Venados <-read.csv(file.choose())
> wilcox.test(Venados$Antes, Venados$Despues, paired=TRUE)
```

Wilcoxon signed rank test with continuity correction

```
data: Venados$Antes and Venados$Despues
V = 0, p-value = 0.005889
alternative hypothesis: true location shift is not equal to 0
```

Aplicaciones de Estadística Básica

Dado $\alpha = 0.05$, el valor de “p” es menor que 0.05, por lo que rechazamos la hipótesis nula y concluimos que hay diferencia entre los pesos de los venados de un momento al otro.

Análisis de varianza para un factor

El análisis de varianza (ANDEVA – ANOVA por sus siglas en inglés-) compara más de dos conjuntos (grupos) de datos de una misma variable y genera un valor de significancia que nos sirve para decidir si los conjuntos de datos son estadísticamente semejantes o diferentes. Las hipótesis en el ANDEVA son:

$$H_0: \mu_1 = \mu_2 = \mu_3 \dots \mu_p$$

H_1 = Al menos una media no es igual.

Donde,

μ = Media.

Para ejemplificar la aplicación del ANDVA se utilizarán los datos de humedad de suelo (%) que hemos venido utilizando, más una columna extra que corresponde a la información del sitio 3. Esta vez los datos se presentarán agrupados en filas, donde la primera columna corresponde a los sitios y la segunda a los valores de humedad de cada sitio. Diríjase al anexo 1 para explorar la tabla de datos completa. La información la importaremos a R y la guardaremos en una variable llamada “HS”. Adicionalmente exploraremos los datos en R, utilizando tres funciones: la función “head()” despliega los primeros seis valores de la tabla de datos, la función “tail()” muestra los últimos seis valores y la función “unique()” permite visualizar los valores únicos en la variable categórica “Sitios”:

```
> HS <- read.csv(file.choose())
> head(HS)
  Sitios  HS
1 Sitio1 85.4
2 Sitio1 91.2
3 Sitio1 93.4
4 Sitio1 84.3
5 Sitio1 86.5
6 Sitio1 98.2
> tail(HS)
  Sitios  HS
25 Sitio3 85.8
26 Sitio3 97.5
```

```
27 Sitio3 93.6
28 Sitio3 76.5
29 Sitio3 95.4
30 Sitio3 82.5
> unique(HS$Sitios)
[1] Sitio1 Sitio2 Sitio3
Levels: Sitio1 Sitio2 Sitio3
```

Aplicaremos el ANDEVA con el uso de la función “aov()” y con el operador “~” que representa el argumento “en función de”, de tal forma que $X \sim Y$ se interpreta como “X en función de Y”, el resultado del análisis lo guardaremos en una variable que se llamará “ANDEVA_HS”:

```
> ANDEVA_HS <- aov(HS$HS ~ HS$Sitios)
> summary(ANDEVA_HS)
              Df Sum Sq Mean Sq F value    Pr(>F)
HS$Sitios      2  484.1   242.06    6.019 0.00689 **
Residuals     27 1085.9    40.22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La función “summary()” la utilizamos para visualizar la tabla de resultados del ANDEVA y como argumento para la función, se escribió el nombre de la variable en donde se guardó el análisis (ANDEVA_HS). Dado $\alpha = 0.05$, el valor de “p” ($\text{Pr}(>F) = 0.00689$) es menor de 0.05, por lo que no tenemos evidencias para afirmar que H_0 es falsa, de tal forma se confirma que existen diferencias significativas entre los datos de cada sitio.

Si deseamos indicarle al programa la condición de la igualdad de varianza, debemos utilizar la función “oneway.test()” en lugar de “aov()” y establecer en el argumento “var.equal=TRUE” si consideramos que las varianzas son semejantes y en caso contrario no escribimos el argumento o escribimos “var.equal=FALSE”:

```
> oneway.test(HR$HR ~ HR$Sitios, var.equal=TRUE)
```

One-way analysis of means

```
data: HR$HR and HR$Sitios
F = 6.0187, num df = 2, denom df = 27, p-value = 0.006893
```

Pruebas de comparaciones múltiples

Anteriormente aplicamos análisis de varianza a tres conjuntos de datos para determinar diferencias significativas entre ellos, el ANDEVA nos demostró que existen dichas diferencias, pero no nos ha determinado cuál o cuáles de los conjuntos de datos están promoviendo las diferencias, para esto último se utilizan las pruebas de comparaciones múltiples. Estas pruebas comparan los grupos de una forma pareada y determinan las parejas de conjuntos con diferencias significativas. Hay diferentes tipos de pruebas de comparación múltiples, cada cual con sus ventajas y desventajas, entre ellas: Fisher LSD, Bonferroni, Tukey, Duncan, Scott Knott, Scheffé, entre otras.

Con el fin de ilustrar el uso de estas pruebas, aplicaremos la prueba de comparación múltiple de Tukey a los datos de humedad de suelo. R cuenta con la función “TukeyHSD()” para lograr este cometido, la cual se puede encontrar entre las funciones de R básico. Para otras pruebas similares tendríamos que instalar paquetes específicos, por ejemplo el paquete “agricolae” posee funciones para aplicar pruebas de comparación múltiple de “LSD.test()” y “duncan.test()”. Utilizamos los resultados del ANDEVA aplicado anteriormente y guardado en la variable “ANDEVA_HS” como argumento de la función “TukeyHSD()”:

```
> TukeyHSD(ANDEVA_HS)
Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = HS$HS ~ HS$Sitios)

$`HS$Sitios`
      diff      lwr      upr      p adj
Sitio2-Sitio1 -8.85 -15.881894 -1.818106 0.0115051
Sitio3-Sitio1 -0.70  -7.731894   6.331894 0.9670163
Sitio3-Sitio2  8.15   1.118106  15.181894 0.0206583
```

Dado $\alpha = 0.05$, el valor de “p” (p adj) es menor de 0.05 al comparar los grupos “Sitio2-Sitio1” y “Sitio3-Sitio2” por lo que concluimos que hay diferencias significativas entre las medias de este par de comparaciones. Adicionalmente notamos que el valor de “p” es mayor de 0.05, al comparar los grupos “Sitio3-Sitio1”, de tal forma que concluimos que no hay diferencia en esta comparación.

Con lo anterior deducimos que el “Sitio2” es el que está marcando las diferencias. Ahora calcularemos las medias para los tres grupos a fin de determinar porqué el “Sitio2” es diferente. Esto lo logramos utilizando la función “tapply()” y tres argumentos, los

valores numéricos, los valores categóricos y el argumento “FUN=mean” que le indica al programa el cálculo de la media para cada grupo de datos:

```
> tapply(HS$HS, HS$Sitios, FUN=mean)
Sitio1 Sitio2 Sitio3
88.98  80.13  88.28
```

Con el cálculo de las medias observamos claramente que el “Sitio 2” es diferente a los otros dos, porque tiene el menor valor promedio (80.13%) de humedad de suelo.

Además de la antes discutida, también se puede aplicar la función “pairwise.wilcox.test()”, en esta se puede personalizar el método para ajustar el valor de “p” con el argumento “p.adj=”:

```
> pairwise.t.test(HR$HR, HR$Sitios, paired=FALSE, p.adj="holm")
```

```
Pairwise comparisons using t tests with pooled SD

data:  HR$HR and HR$Sitios

      Sitio1 Sitio2
Sitio2 0.013  -
Sitio3 0.807  0.016

P value adjustment method: holm
```

Los resultados nos lo presentan en formato de matriz, donde el valor del “p” se obtiene haciendo coincidir la columna y la fila de las categorías (los sitios), por ejemplo el valor de “p” de la comparación “Sitio1-Sitio2” es 0.013 y para la comparación “Sitio2-Sitio3” es 0.016. Notemos que aunque los valores de “p” resultantes no son iguales a los determinados con la función “TukeyHSD()”, hemos llegado a las mismas conclusiones.

Otras opciones de métodos que podemos utilizar para el ajuste del valor de “p” además de “holm”, son: “hochberg”, “hommel”, “bonferroni”, “BH”, “BY”, “fdr”; cuando no deseamos incluir algún método de ajuste escribimos “none” (ninguno en español) en el argumento “p.adj=”.

Si los datos no cumplen con el supuesto de normalidad, podemos utilizar la función “pairwise.wilcox.test()” con los mismos argumentos expresados para “pairwise.t.test()”. Si las medidas son repetidas utilizamos el argumento “paired=TRUE”. Si queremos ajustar las colas utilizamos los argumentos: “alternative=”two.sided”” (por defecto), “alternative=”less”” o “alternative=”greater”” según sea el objetivo de la comparación.

Análisis de varianza no paramétrica para un factor (Kruskal-Wallis)

La prueba de Kruskal-Wallis es el equivalente no paramétrico al análisis de varianza, la ejecutamos con la función “`kruskal.test()`”. Su uso lo ejemplificaremos con los mismos datos de humedad de suelo utilizados para ejemplificar las ANDEVAS, aclarando que esta prueba solamente se aplica después que se haya confirmado que los datos no cumplen con el supuesto de normalidad y que una transformación de datos no es suficiente:

```
> kruskal.test(HS$HS ~ HS$Sitios)

Kruskal-Wallis rank sum test

data:  HS$HS by HS$Sitios
Kruskal-Wallis chi-squared = 11.249, df = 2, p-value = 0.003609
```

Dado $\alpha = 0.05$, el valor de “p” (0.003609) es menor de 0.05, por lo que tenemos evidencias para rechazar H_0 , de tal forma que confirmamos que existen diferencias significativas de las medias de humedad de suelo comparadas entre los sitios.

Análisis de varianza para un factor y medidas repetidas

Se utiliza para comparar las medias de más de dos conjuntos de datos pareados. Las hipótesis a poner a prueba son:

H_0 : No hay cambios en las medidas.

H_1 : Al menos una medida es diferente.

Para ejemplificar, haremos uso de los datos de peso de un grupo de 10 venados que utilizamos anteriormente, agregándole un momento más para un total de tres. El objetivo de aplicar la prueba es que exploremos si el peso de los venados “durante los tres momentos” cambió significativamente. A continuación se muestra una parte de los datos y los datos completos se presentan en el anexo 2:

```
> Venados <- read.csv(file.choose())
> head(Venados)
  ID Momentos  Peso
1  1  Momento1  78.5
2  2  Momento1 123.2
3  3  Momento1 150.8
4  4  Momento1 121.3
5  5  Momento1  79.8
```



```
6 6 Momento1 98.5
> tail(Venados)
      ID Momentos  Peso
25 5 Momento3 89.6
26 6 Momento3 128.7
27 7 Momento3 105.3
28 8 Momento3 110.4
29 9 Momento3 162.3
30 10 Momento3 109.1
> unique(Venados$Momentos)
[1] Momento1 Momento2 Momento3
Levels: Momento1 Momento2 Momento3
```

Notemos que la columna denominada “ID” representa el número de observaciones (10) por conjunto de datos. El análisis lo llevaremos a cabo utilizando las funciones “lme()” abreviado de “Linear Mixed-Effects Models” (modelos de efecto lineal mixto) y “anova()”. Para esto, tenemos que instalar y hacer disponible el paquete llamado “nlme”:

```
> install.packages("nlme")
> library("nlme")
```

El análisis lo ejecutaremos en dos pasos, primero creamos el modelo lineal con “lme()” y lo guardamos en la variable “ANDEVA_Repet” y luego aplicamos la función “anova()” a dicho resultado:

```
> ANDEVA_Repet <- lme(Peso ~ Momentos, random= ~1|ID/Momentos,
data=Venados)
> anova(ANDEVA_Repet)
      numDF denDF    F-value p-value
(Intercept)      1     18 199.13343  <.0001
Momentos         2     18  28.65142  <.0001
```

El argumento “Peso ~ Momentos” establece la comparación de los pesos en función de los momentos; el argumento “random= ~1|ID/Momentos” determina el modelo (en el argumento “random=” se establece el modelo, la primera parte (~1|) es recomendable copiarla igual, “ID” es el número de observaciones por conjunto de datos y “Momentos” es el nombre que define a cada momento. “ID” y “Momentos” son también conocidos como factores y el argumento “data=Venados” le indica al programa en que variable están guardados los datos de “Peso” y “Momentos”.

Aplicaciones de Estadística Básica

Dado $\alpha = 0.05$, el valor de “p” (<0.0001) es menor de 0.05, por lo que tenemos evidencias para rechazar la H_0 , de tal forma que confirmamos que existen diferencias significativas del peso de los venados entre los tres momentos.

Con el ANDEVA de medidas repetidas, solamente hemos determinado la existencia de diferencias significativas entre los conjuntos de datos (Momentos), pero no se ha determinado cuáles conjuntos están haciendo las diferencias, para determinarlo aplicaremos una prueba de comparación múltiple. La ejemplificación la realizaremos aplicando la prueba de Tukey, utilizando el paquete “multcomp” en donde aplicará la función “glht()” o hipótesis lineal general y comparación múltiple para modelos paramétricos:

```
> install.packages("multcomp")
> library("multcomp")
> summary(glht(ANDEVA_Repet, linfct=mcp(Momentos="Tukey")))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

```
Fit: lme.formula(fixed = Peso ~ Momentos, data = Venados, random = ~1 |
      ID/Momentos)
```

Linear Hypotheses:

	Estimate	Std. Error	z value	Pr(> z)	
Momento2 - Momento1 == 0	5.440	1.907	2.853	0.0121	*
Momento3 - Momento1 == 0	14.300	1.907	7.499	<0.001	***
Momento3 - Momento2 == 0	8.860	1.907	4.646	<0.001	***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Con los comandos anteriores, ejecutamos la prueba y a la vez indicamos al programa que presentará las tablas de resultados, por medio de la función “summary()”. Los datos utilizados los teníamos previamente almacenados en la variable “ANDEVA_Repet”. El argumento “linfct=mcp(Momentos=“Tukey”)” establece la prueba de multicomparación a ejecutar. En la función “glht()” establecemos los argumentos: “alternative=“two.sided”” (por defecto), “alternative=“less”” o “alternative=“greater””, según sea el objetivo de la comparación.

Dado $\alpha = 0.05$, los valores de “p” ($\Pr(>|z|)$) son todos menores a 0.05, por lo que tenemos evidencias para rechazar las tres H_0 y concluir que los pesos de los venados en los tres momentos son significativamente diferentes, en especial entre “Momento 3 – Momento 1” y “Momento 3 – Momento 2” (***), siendo las diferencias también significativas entre “Momento 2 – Momento 1” (*) pero menos que las anteriores.

Análisis de varianza no paramétrico para un factor y medidas repetidas (Prueba de Friedman)

Dentro de la familia de análisis no paramétricos contamos con una prueba para determinar diferencias entre medias de medidas repetidas. La prueba asume las mismas hipótesis que la equivalente paramétrica. Para ejecutar la prueba no paramétrica es necesario que hayamos confirmado que los datos no cumplen con el supuesto de normalidad y que la transformación no es una opción.

Aplicaremos la prueba con el uso de la función “`friedman.test()`” utilizando los datos de peso de los venados que anteriormente empleamos. Es necesario que estos datos contengan una variable para definir el número de observaciones por grupo (factor), en nuestro caso le hemos asignado el nombre de “ID” a esa variable. El comando quedaría estructurado de la siguiente manera:

```
> friedman.test(Venados$Peso, Venados$Momentos, Venados$ID)
```

```
Friedman rank sum test
```

```
data: Venados$Peso, Venados$Momento and Venados$ID  
Friedman chi-squared = 20, df = 2, p-value = 4.54e-05
```

Dado $\alpha = 0.05$, el valor de “p” (p-value) es menor a 0.05, por lo que tenemos evidencias para rechazar H_0 y por lo tanto concluimos que hay diferencias significativas entre los pesos medidos en cada momento.

Análisis de varianza para dos factores

El análisis de varianza para dos factores es la comparación de medias entre más de dos conjuntos de datos, los cuales están agrupados por dos factores, el factor puede estar dividido en diferentes niveles o variables categóricas. El análisis de varianza para dos factores asume las siguientes hipótesis:

Aplicaciones de Estadística Básica

H_0 =No hay diferencias significativas entre las medias de los niveles del factor 1.

H_1 =Al menos una media de uno de los niveles dentro del factor 1 es significativamente diferente.

H_0 =No hay diferencias significativas entre las medias de los niveles del factor 2.

H_1 =Al menos una media de uno de los niveles dentro del factor 2 es significativamente diferente.

H_0 =No hay interacciones significativas entre ambos factores.

H_1 =Hay interacciones significativas entre ambos factores.

ANDEVA de dos factores con replicación

El análisis de varianza de dos factores con replicación, es el que se aplica a diseños en los cuales los niveles de dicho factor tienen muestras replicadas. Para realizar el análisis en R se utiliza la función “aov()” y en los argumentos se especifica el modelo. A diferencia de MS Excel, en R se utiliza el arreglo en filas para aplicar la función. Si “dato” es la columna que contiene los valores numéricos, “x” es el primer factor y “y” es el segundo factor, los argumentos en la función “aov()” los arreglaríamos de la siguiente manera: aov(dato ~ x + y + x*y) en donde se calcularán el valor de la probabilidad para “x”, “y” y la interacción entre factores.

Para ejemplificar la aplicación de este análisis, utilizaremos unos datos de concentración de Oxígeno disuelto (OD) (ppm) en el agua, a lo largo de un río principal en una microcuenca. Los muestreos se realizaron en las tres partes de la microcuenca, las cuales son: parte alta, parte media y parte baja; adicionalmente, en cada parte se selecciona dos usos de suelo, estos son el uso bosque y el uso agrícola, y en cada uno de estos usos se establecieron cinco puntos de muestreo (réplicas). Se pretende determinar si existen diferencias entre las concentraciones de OD entre las partes de las microcuencas (factor 1) y los usos del suelo (factor 2) y entre sus interacciones. Primeramente importamos los datos a R y los guardamos en la variable “ANDEVA2”, a continuación se muestra parte de la información, los datos completos se encuentran en anexo 3:

```
> ANDEVA2 <-read.csv(file.choose())
> head(ANDEVA2)
  Toposec Ecosist OD
1    Alta  Bosque 4.5
2    Alta  Bosque 5.3
3    Alta  Bosque 4.3
4    Alta  Bosque 4.8
5    Alta  Bosque 4.7
6    Alta Agrícola 3.7
> tail(ANDEVA2)
```

```
Toposec Ecosist OD
25 Baja Bosque 3.1
26 Baja Agrícola 2.3
27 Baja Agrícola 3.2
28 Baja Agrícola 1.5
29 Baja Agrícola 1.6
30 Baja Agrícola 2.3
> unique(ANDEVA2$Toposec)
[1] Alta Media Baja
Levels: Alta Baja Media
> unique(ANDEVA2$Ecosist)
[1] Bosque Agrícola
Levels: Agrícola Bosque
```

Seguidamente aplicamos la prueba y guardamos los resultados en la variable “ANDEVA2Result”. La tabla de resultados la obtenemos utilizando la función “summary()”:

```
> ANDEVA2Result <-aov(ANDEVA2$pH ~ ANDEVA2$Toposec +
  ANDEVA2$Ecosist + ANDEVA2$Toposec*ANDEVA2$Ecosist)
> summary(ANDEVA2Result)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
ANDEVA2\$Toposec	2	19.51	9.756	20.95	5.44e-06	***
ANDEVA2\$Ecosist	1	4.80	4.800	10.31	0.00374	**
ANDEVA2\$Toposec:ANDEVA2\$Ecosist	2	0.00	0.000	0.00	1.00000	
Residuals	24	11.18	0.466			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Dado $\alpha = 0.05$, los valores de “p” ($\text{Pr}(>F)$) para los factores toposecuencia (ANDEVA2\$Toposec) y Ecosistemas (ANDEVA2\$Ecosist) son menores a 0.05, por lo que tenemos evidencias para rechazar la H_0 , de tal forma que confirmamos que existen diferencias significativas al comparar los datos entre los niveles de cada factor; sin embargo, el valor de “p” de la interacción fue mayor a 0.05, por lo que concluimos que no hay interacciones entre los factores.

Si el comando para ejecutar el análisis se nos hace muy largo con el uso del símbolo de dólar (\$), podemos utilizar otro formato de escritura donde la fuente de datos (variable que contiene de datos) no se llama con dicho símbolo, sino que con el argumento “data=”, los resultados son absolutamente los mismos:

```
> ANDEVA2Result <-aov(pH ~ Toposec + Ecosist + Toposec*Ecosist,
  data=ANDEVA2)
```

ANDEVA de dos factores sin replicación

El análisis de varianza de dos factores sin replicación es el que se aplica a diseños, en los cuales los niveles de uno de los factores no tienen muestras replicadas. Para realizar el análisis en R se utiliza la función “aov()” y en los argumentos se especifica el modelo. A diferencia de MS Excel, en R utilizamos el arreglo de datos por filas (Figura 84) para aplicar la función. Si “dato” es la columna que contiene los valores numéricos, “x” es el primer factor y “y” es el segundo factor, los argumentos en la función “aov()” se arreglarían de la siguiente manera: aov(dato ~ x + y).

Para ejemplificar la aplicación de este análisis utilizaremos unos datos de Oxígeno Disuelto tomado en orilla de diferentes ríos en las tres partes de una microcuenca: parte alta, parte media y parte baja (Factor 1); se realizó un muestreo en cada uno de cinco ríos (Factor 2) en cada parte de la microcuenca. Primeramente importamos los datos a R y los guardamos en la variable “ANDEVA2NoRep”, a continuación se muestra parte de la información, los datos completos se encuentran en anexo 4:

```
> ANDEVA2NoRep <-read.csv(file.choose())
```

```
> head(ANDEVA2NoRep)
```

	Toposec	Ríos	OD
1	Alta	Rio Grande	5.6
2	Alta	Rio Escondido	4.5
3	Alta	Rio El Salto	4.2
4	Alta	Rio Alegre	5.3
5	Alta	Rio San Luis	2.1
6	Media	Rio Grande	4.3

```
> tail(ANDEVA2NoRep)
```

	Toposec	Ríos	OD
10	Media	Rio San Luis	3.4
11	Baja	Rio Grande	3.2
12	Baja	Rio Escondido	3.6
13	Baja	Rio El Salto	3.2
14	Baja	Rio Alegre	4.1
15	Baja	Rio San Luis	2.9

Notemos que el factor 2 no se replica para los niveles del factor 1. Seguidamente aplicamos la prueba y guardamos los resultados en la variable “ANDEVA2NoRepResult”. La tabla de resultados la obtenemos utilizando la función “summary()”:

```
> ANDEVA2NoRepResult <-aov(ANDEVA2NoRep$OD ~ ANDEVA2NoRep$Toposec + ANDEVA2NoRep$Ríos)
```

```
> summary (ANDEVA2NoRepResult)
              Df Sum Sq Mean Sq F value Pr(>F)
ANDEVA2NoRep$Toposec  2  2.249   1.1247    2.718 0.1257
ANDEVA2NoRep$Ríos    4  5.689   1.4223    3.437 0.0646 .
Residuals            8  3.311   0.4138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado $\alpha = 0.05$, los valores de “p” ($\Pr(>F)$) para los factores Toposecuencia (ANDEVA2NoRep\$Toposec) y Ríos (ANDEVA2NoRep\$Ríos) son mayores a 0.05, por lo que en ambos casos no encontramos evidencias para rechazar la H_0 , y confirmamos que no hay diferencias significativas al comparar las medias entre los niveles de cada factor.

Si el comando para ejecutar el análisis se nos hace muy largo con el uso del símbolo de dólar (\$), podemos utilizar otro formato de escritura donde la fuente de dato (variable que contiene los datos) no se llama con dicho símbolo, sino que con el argumento “data=”, los resultados son absolutamente los mismos:

```
> ANDEVA2NoRepResult <- aov (pH ~ Toposec + Ríos, data=ANDEVA2NoRep)
```

Relaciones entre variables

Dos variables están relacionadas sí y solamente si los valores de una incrementan o disminuyen en función de los valores de la otra, es decir los valores de la variable “X”, por ejemplo, incrementan o disminuyen a la vez que los valores de la variable “Y” también incrementan o disminuyen. La búsqueda de relaciones entre variables, es una tarea común en estadística básica y paso importante para otros análisis.

Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson explora la relación entre conjuntos de datos de dos variables, a fin de determinar cómo se comporta una en relación a la otra. Con este coeficiente se calcula la significancia, fuerza y dirección de la relación. Las relaciones también las podemos explorar gráficamente (Figura 50) y de esa forma obtenemos información general sobre el tipo de relación. Las hipótesis de la prueba son:

H_0 =No hay correlación significativa.

H_1 =Hay correlación significativa.

Una correlación positiva indica que los valores de una variables incrementan al incrementar los valores de la otra o disminuyen al disminuir los valores de la otra; una corre-

Aplicaciones de Estadística Básica

lación negativa significa que los valores de una variable se reducen al incrementar los valores de la otra o viceversa; cuando no hay correlación las variables se comportan de forma independiente.

Podemos visualizar la relación con la función “plot()” y calcular el coeficiente de correlación de Pearson en R con el uso de la función “cor.test()”. Para ejemplificar haremos uso de datos de elevación y temperatura utilizados para ejemplificar la misma prueba en MS Excel:

```
> Elev_Temp <- read.csv(file.choose())
> Elev_Temp
  Elevacion Temp
1    198.7  12.3
2    182.0  13.2
3    183.3  16.0
4    154.3  16.3
5    163.3  18.5
6    132.2  22.5
7    140.2  18.9
8    120.9  24.7
```

Para crear el gráfico, colocamos la variable elevación en el eje X y la variable temperatura en el eje Y (Figura 89):

```
> plot(Elev_Temp$Elevacion, Elev_Temp$Temp)
```

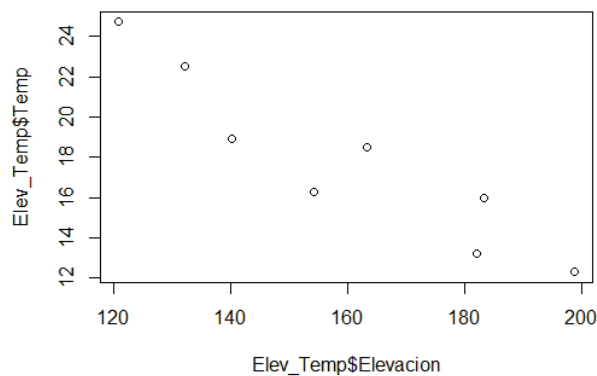


Figura 89. Gráfico de puntos mostrando la correlación entre las variables elevación y temperatura.

Observamos que el gráfico muestra una correlación negativa al compararlo con los ejemplos ilustrados en la figura 50. La personalización del gráfico no se abordará en este instante, sino en el tema correspondiente a “Opciones gráficas”.

A continuación se aplica la función “cor.test()”, escribiendo dentro de la función los argumentos que indican los valores de cada variable y el tipo de método a usar, en este caso el Coeficiente de Correlación de Pearson:

```
> cor.test(Elev_Temp$Elevacion, Elev_Temp$Temp, method="pearson")

Pearson's product-moment correlation

data: Elev_Temp$Elevacion and Elev_Temp$Temp
t = -6.4166, df = 6, p-value = 0.0006763
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9882893 -0.6719099
sample estimates:
      cor
-0.9342413
```

La tabla de resultado nos presenta varios valores calculados, entre ellos el coeficiente de correlación (-0.9342413), del cual deducimos que la relación es negativa (por el signo negativo) y muy fuerte pues el valor se acerca a 1.0; además, calcula el valor de “p” (0.0006763) en cual, dado un $\alpha = 0.05$, este es mucho menos a 0.05, por lo que tenemos evidencias para rechazar H_0 , y concluimos que la relación es significativa. El resultado que provee la función también ofrece un intervalo de confianza del 95%.

Correlación no paramétrica (Coeficiente de Correlación de Spearman)

El Coeficiente de Correlación de Spearman es la alternativa no paramétrica cuando hemos probado que los valores de las variables no se distribuyen de forma normal. Este coeficiente tiene la misma función que el Coeficiente de Correlación de Pearson, solamente que el algoritmo para su ejecución es diferente a fin de dar salida a un cálculo no paramétrico.

Este coeficiente se calcula con la función “cor.test()” y cambiando el argumento de “method=“pearson”” a “method=“spearman””. Cambiando el argumento a “method=“kendall”” se obtiene la correlación con el procedimiento de Kendall. Utilizaremos los datos de elevación y temperatura para ejemplificar:

Aplicaciones de Estadística Básica

```
> cor.test(Elev_Temp$Elevacion, Elev_Temp$Temp, method="spearman")

Spearman's rank correlation rho

data: Elev_Temp$Elevacion and Elev_Temp$Temp
S = 164, p-value = 0.001141
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.952381
```

Confirmamos que la relación es negativa, muy fuerte (el valor del coeficiente se acerca a 1.0) y significativa (dado un $\alpha = 0.05$, el valor de “p” es mucho menos a 0.05, por lo que hay evidencias para rechazar H_0).

Regresión lineal simple

Después de confirmar que dos variables están correlacionadas significativamente, por lo general los investigadores están interesados en predecir los valores de una variable (variable dependiente) en función de la otra (variable independiente). Hay muchas formas de determinar cuál de dos variables es dependiente o independiente, lo más intuitivo es indagar que variable está influenciada por la otra. Por ejemplo: la temperatura y la elevación (sobre el nivel del mar), en general, tienen correlación negativa, o sea que al aumentar los valores de una, disminuyen los valores de la otra, es decir al incrementar la elevación se reduce la temperatura (se torna más frío).

Para definir cuál de las dos variables es la dependiente y cual la independiente se puede recurrir al siguiente razonamiento: “al aumentar los valores de la variable elevación se esperaría (evidentemente) que la temperatura se reduzca, de igual forma al disminuir la elevación se esperaría que la temperatura incrementara; sin embargo, al aumentar o reducir los valores de la variable temperatura no se esperaría ningún cambio en la variable elevación”. De lo anterior se deduce de este dueto de variables que la elevación es una variable influyente y la temperatura es una variable influida, por lo que les llamamos independiente y dependiente respectivamente. Para que utilicemos el análisis de regresión, es necesario que definamos la variable dependiente e independiente, el no definir las es probable que no tenga considerable impacto en los cálculos, pero sí en la interpretación.

En R, los cálculos de regresión se logran con la función “lm()” cuyos resultados se guardan en una variable y se despliegan con el uso de la función “summary()”. Para

ejemplificar la aplicación se hará uso de los datos de la tabla de elevación y temperatura usada anteriormente, los cuales se guardarán en una variable a la que llamaremos “Elev_Temp”:

```
> Elev_Temp <- read.csv(file.choose())
> Elev_Temp
  Elevacion Temp
1    198.7  12.3
2    182.0  13.2
3    183.3  16.0
4    154.3  16.3
5    163.3  18.5
6    132.2  22.5
7    140.2  18.9
8    120.9  24.7
```

Si “Y” nos representa a una variable dependiente y “X” a una variable independiente, los argumentos dentro de la función los expresaríamos como “lm(Y ~ X)”, o sea, “Y” en función de “X”. Dado “Temp” la variable dependiente y “Elevacion” la variable independiente, la función quedaría definida de la siguiente manera:

```
> Reg <- lm(Elev_Temp$Temp ~ Elev_Temp$Elevacion)
> summary(Reg)
```

Call:

```
lm(formula = Temp_Hum$Temp ~ Temp_Hum$Elevacion)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.2388	-1.3965	0.4884	1.2777	1.6932

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.05583	3.67094	11.184	3.05e-05 ***
Temp_Hum\$Elevacion	-0.14593	0.02274	-6.417	0.000676 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.649 on 6 degrees of freedom

Multiple R-squared: 0.8728, Adjusted R-squared: 0.8516

F-statistic: 41.17 on 1 and 6 DF, p-value: 0.0006763

Aplicaciones de Estadística Básica

La función despliega los resultados de varios análisis estadísticos, en el caso particular de esta publicación solamente fijaremos la atención en tres de ellos: 1. El valor de p de los coeficientes ($\Pr(>|t|)$); 2. El valor de coeficiente de determinación (R^2) (R-squared) y 3. La probabilidad (p-value). Dado $\alpha = 0.05$, el valor de “ p ” de los coeficientes nos indican que son significativos y pueden ser usados en el modelo; el coeficiente de determinación (0.8728) se acerca a 1.0 denotándonos una correlación fuerte, cuando R^2 se acerca a 1 nos indica que la mayoría de la información se incluyó en el modelo y que el modelo es certero y representativo, cuando R^2 se acerca a 0 significa que muy poca información se incluyó en el modelo y que dicho modelo no estaría representando la verdadera relación entre las dos variables; y el valor de “ p ” (0.0006763) nos determina la significancia del modelo de regresión.

La relación la podemos visualizar al agregar una línea de regresión al gráfico de punto mostrado en la figura 90, para ello se utilizará la función “`abline()`” y el modelo de regresión con la función “`lm()`”:

```
> plot(Elev_Temp$Elevacion, Elev_Temp$Temp)
> abline(lm(Elev_Temp$Temp ~ Elev_Temp$Elevacion))
```

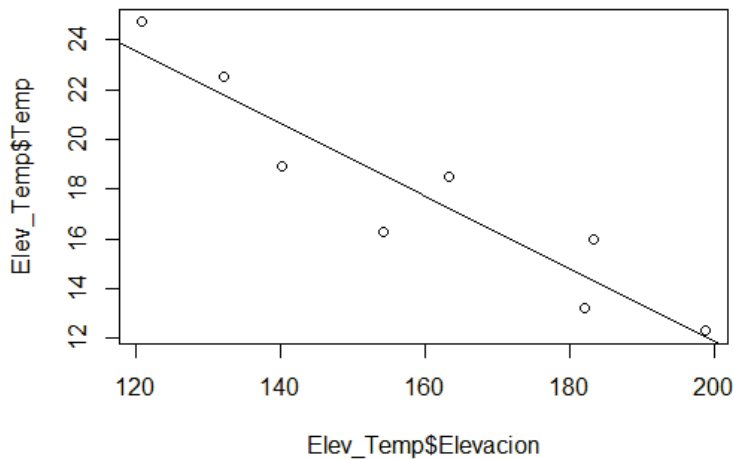


Figura 90. Gráfico de puntos mostrando la correlación y línea de regresión entre las variables elevación y temperatura.

El gráfico anterior nos muestra la correlación negativa que existe entre las variables elevación y tempera, la personalización del gráfico no lo abordaremos en este instante, sino en el tema correspondiente a “Opciones gráficas”.

En R al igual que en MS Excel, podemos visualizar los residuales mediante un gráfico, esto nos permite evaluar el modelo visualmente y determinar los valores atípicos. Para ello, utilizamos la función “plot()” y como argumento se escribe la variable en donde se guardó el modelo de regresión, en nuestro caso la variable “Reg”, además del argumento “which=” con el cual seleccionamos el tipo de gráfico entre las opciones: 1= residuales versus valores ajustados; 2= Gráfico Q-Q; 3= Escala-Localización; 4= Distancia de Cook y 5= Residuos versus leverage (que tan distante están los valores de la variable independiente de una observación en relación a otras observaciones), para este ejemplo utilizaremos la opción 1 (Figura 91 A):

```
> plot(Reg, which=1)
```

Si quisiéramos limpiar un poco el gráfico, hacerlo más sencillo de interpretar y más estético, añadiremos al comando otros argumentos, entre ellos “add.smooth=FALSE” con el cual quitaremos la línea suavizada que se le agrega por defecto al gráfico; además utilizaremos el argumento “caption=NA” a fin de quitar el título principal del gráfico, entonces el comando se visualizaría de la siguiente forma (Figura 91 B):

```
> plot(Reg, which=1, add.smooth=FALSE, caption=NA)
```

La distribución de los puntos en el gráfico nos demuestra que no hay ninguna condición extraña en el modelo (comparada con la figura 56), sin embargo el programa nos resalta los valores extremos (3, 4 y 7) los cuales podrían ser potenciales valores atípicos que pudiesen influir en el modelo.

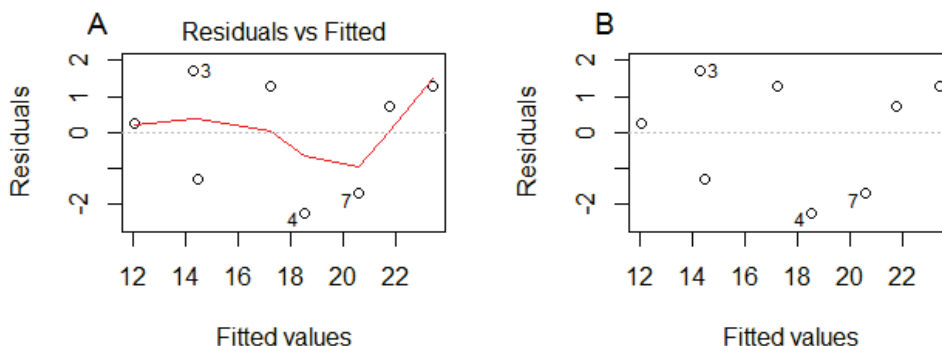


Figura 91. Gráfico de residuales para el modelo de regresión lineal simple. A. Gráfico no personalizado; B. Gráfico personalizados.

Aplicaciones de Estadística Básica

Con el argumento “which=” se pueden seleccionar cinco opciones más de gráficos, aparte de el gráfico de residuales presentado anteriormente. Los gráficos son útiles para evaluar el modelo, sin embargo en este escrito no se ampliará en la explicación de cada uno, por lo que se le recomienda al lector utilizar referencias adicionales. A continuación se mostrarán todos los gráficos a los que tenemos acceso con el argumento “which=” (Figura 92):

```
> plot(Reg, which=1) #Ver Figura 92 A
> plot(Reg, which=2) #Ver Figura 92 B
> plot(Reg, which=3) #Ver Figura 92 C
> plot(Reg, which=4) #Ver Figura 92 D
> plot(Reg, which=5) #Ver Figura 92 E
> plot(Reg, which=6) #Ver Figura 92 F
```

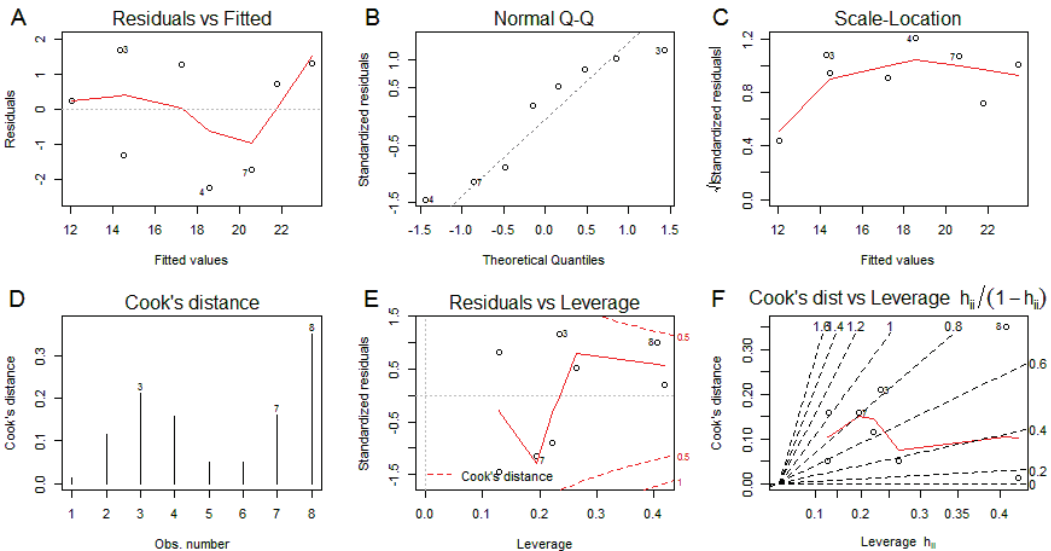


Figura 92. Diferentes opciones gráficas que se despliegan con el uso de del argumento “which=” dentro de la función “plot()” para el modelo de regresión. A. Residuales versus valores ajustados; B. Gráfico Q-Q; C. Residuales estandarizados versus valores ajustados; D. Distancia de Cook; E. Residuales estandarizados versus valores leverage (que tan distante están los valores de la variable independiente de una observación en relación a otras observaciones); F. Distancia de Cook versus valores leverage.

La presencia de datos atípicos se puede evaluar con los gráficos anteriores y se puede confirmar con una prueba específica que para tales fines nos ofrece R en el paquete “car”. Para ejecutar la prueba se hace uso de la función “outlierTest()”, con esta se puede confirmar la presencia y a la vez seleccionar los datos atípicos. Instalaremos el paquete con sus dependencias, o sea con otros paquetes que “car” necesita para su funcionamiento, por ello agregamos el argumento “dependencies=TRUE” en la función “install.package()”. Después de instalar y hacer disponible el paquete “car”, corremos la prueba con la función y el único argumento es el modelo de regresión guardado en la variable “Reg”:

```
> install.package("car", dependencies=TRUE)
> library("car")
> outlierTest(Reg)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferonni p
4 -1.651578          0.15953          NA
```

La función ha detectado a la observación 4 como potencial datos atípicos, en el modelo de regresión simple, sin embargo este no es significativo. Con este conjunto de datos en particular no se calculó la p ajustada por el método de Bonferonni.

Si se desea determinar las observaciones que influyen en el modelo de regresión, se puede utilizar la función “influence.measures()” incluida entre las funciones de R básico, con esta se obtienen diferentes medidas y pruebas para alcanzar tales fines. La función señala a las observaciones que más influyen en el modelo con un asterisco (*). Más información sobre la función se puede determinar escribiendo y ejecutando el argumento “?influence.measures” en la consola.

```
> influence.measures(Reg)
Influence measures of
      lm(formula = Elev_Temp$Temp ~ Elev_Temp$Elevacion) :

      dfb.1_ dfb.E_T.  dffit cov.r cook.d  hat inf
1 -0.11026   0.1248   0.149 2.449 0.0132 0.419  *
2  0.24977  -0.3093  -0.467 1.394 0.1137 0.222
3 -0.37596   0.4601   0.674 1.116 0.2101 0.234
4 -0.22147   0.1236  -0.638 0.693 0.1581 0.130
5  0.00219   0.0467   0.307 1.296 0.0502 0.128
6  0.24179  -0.2126   0.292 1.787 0.0490 0.265
7 -0.41840   0.3487  -0.582 1.091 0.1590 0.195
8  0.76429  -0.6991   0.840 1.667 0.3511 0.406
```

Aplicaciones de Estadística Básica

Concluimos que de las ocho observaciones que forman parte del modelo, solamente la número uno (elevación= 198.7, temperatura= 12.3) influencia significativamente en el modelo.

Llegando un poco más allá con las pruebas para evaluar el modelo, en R también tenemos disponible una prueba para determinar la autocorrelación de los residuales. La autocorrelación es la dependencia de una variable consigo misma en una escala de tiempo, que distorsiona los estadísticos del modelo, la presencia de autocorrelación indica la carencia de una variable predictora útil. Las hipótesis de la prueba son las siguientes:

H_0 =No existe autocorrelación significativa.

H_1 =Existe autocorrelación significativa.

La prueba la ejecutamos con la función “dwtest()”, que se encuentra en el paquete llamado “lmtest”, de tal forma que instalamos el paquete y corremos la función con el modelo de regresión como argumento:

```
> install.packages("lmtest")  
> library("lmtest")  
> dwtest(Reg)
```

Durbin-Watson test

```
data: Reg  
DW = 3.3222, p-value = 0.9915  
alternative hypothesis: true autocorrelation is greater than 0
```

Dado un $\alpha = 0.05$, el valor de “p” es mucho mayor, por lo que no hay evidencias para rechazar H_0 , o sea no existe autocorrelación entre los valores residuales.

Para realizar una predicción de la variable dependiente (“Temperatura”) en base a los valores de la variable independiente (“Elevación”), utilizaremos la fórmula:

$$Y = a + b(X)$$

Donde,

Y = Variable dependiente (a predecir)

a = Intercepción

b = Pendiente

X = Variable independiente (predictora)

Si, $a = 41.06$ y $b = -0.15$, la fórmula de predicción quedaría como:

$$Y = 41.06 + (-0.15)(X)$$

Para usar la fórmula, estableceremos que, por ejemplo, se necesitaría predecir la temperatura (Y) cuanto la elevación (X) es 150 m, entonces se sustituye ese valor en la fórmula:

$$Y = 41.06 - 0.15(150) = 18.56 \text{ }^{\circ}\text{C}$$

O sea que a los 150 metros de elevación esperaríamos una temperatura aproximada de 18.56 °C.

Otra prueba para determinar datos atípicos, esta vez no en modelos de regresión, sino en conjuntos de datos, es la que se ejecuta mediante la función “outlier()”, disponible en el paquete “outliers”. La función selecciona los valores extremos tomando como punto de partida la media, de tal forma que se pueden seleccionar los datos atípicos mayores y menores que la media. Para seleccionar los mayores que la media se utiliza la función con solamente la tabla de datos como argumento; para seleccionar los menores se emplea el argumento “opposite= TRUE”. Utilizaremos los datos guardados en la variable “Elev_Temp” para correr la función:

```
> install.packages("outliers")
> library("outliers")
> outlier(Elev_Temp)
Elevacion      Temp
      198.7      24.7
> outlier(Elev_Temp, opposite= TRUE)
Elevacion      Temp
      120.9      12.3
```

Según la función se sospecha que los datos atípicos mayores que la media pueden ser 198.7 de la variable “Elevacion” y 24.7 de la variable “Temp”; y los menores que la media pueden ser 120.9 de la variable “Elevacion” y 12.3 de la variable “Temp”, tomando como referencias que las medias para cada variable las cuales son: 159.4 para la variable “Elevacion” y 17.8 para la variable “Temp”.

Regresión lineal múltiple

El principio de la regresión lineal múltiple es similar al de la regresión lineal simple, solamente que en la múltiple se incluyen más de dos variables, de las cuales una es la

Aplicaciones de Estadística Básica

variable dependiente (o predicha) y el resto son variables independientes (o predictores). Para ejemplificar, utilizaremos una pequeña base de datos formada por cuatro variables, de las cuales la variable “Cober” representa la cobertura (%) de *Anomadon attenuatus* (una especie de musgo epífita) a 20 cm del suelo sobre la base de los árboles y las variables predictores son la temperatura (°C), la humedad relativa del aire (HR) y el diámetro (en cm) del árbol hospedero (DAP). Con el análisis pretendemos predecir la cobertura del musgo con la combinación de valores de las otras variables. A continuación se muestra la pequeña base de datos y se guarda en la variable “MultiVar”:

```
> MultiVar <-read.csv(file.choose())
```

```
> head(MultiVar)
```

	Cober	Temp	HR	DAP
1	81.2	3.2	98.2	80
2	72.1	5.3	91.3	75
3	61.3	6.5	82.6	71
4	52.4	9.6	83.6	63
5	43.2	11.4	75.2	60
6	35.7	13.8	71.7	56
7	26.2	14.2	66.2	52
8	23.2	16.5	63.9	46
9	15.4	18.6	57.1	38
10	8.1	20.6	50.6	30

Utilizaremos la función “lm()” para correr el análisis de regresión lineal múltiple, anexando cada variable con el operador de sumatoria (+), los resultados los guardaremos en la variable que llamaremos “MultiReg”. Seguidamente visualizaremos los resultados con el uso de la función “summary()”:

```
> MultiReg <-lm(MultiVar$Cober ~ MultiVar$Temp + MultiVar$HR +  
MultiVar$DAP)
```

```
> summary(MultiReg)
```

Call:

```
lm(formula = MultiVar$Cober ~ MultiVar$Temp + MultiVar$HR +  
MultiVar$DAP)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.798	-0.003	0.355	1.088	2.146

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	63.1635	41.2489	1.531	0.1766

En Microsoft® Excel y R

```
MultiVar$Temp    -3.8571      1.1719    -3.291    0.0166 *
MultiVar$HR       0.8624      0.3218     2.680    0.0366 *
MultiVar$DAP     -0.6824      0.4138    -1.649    0.1502
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.332 on 6 degrees of freedom
Multiple R-squared:  0.994,    Adjusted R-squared:  0.991
F-statistic: 330.3 on 3 and 6 DF,  p-value: 4.757e-07
```

Otra forma de escribir el comando es: “lm(Coher ~ Temp + HR + DAP, data=MultiVar)”, el lector queda invitado a tratar la forma que más le convenga.

En los resultado observamos tres cosas: 1. El intercepto no fue significativo ($\Pr(>|t|)$) para un $\alpha = 0.05$ y por ende deberemos tener cautela con la interpretación de los resultados de las predicciones, los coeficientes fueron significativos para las variables temperatura (MultiVar\$Temp) y humedad relativa (MultiVar\$HR), no siendo así para la variable DAP (MultiVar\$DAP); 2. El valor del coeficiente de determinación (R^2) fue extremadamente elevado; 3. El valor de p (p-value) es muy pequeño comparado con 0.05, por lo que se concluye que el modelo es significativo.

Hemos observado que la variable “DAP” no está contribuyendo al modelo y es candidata a ser retirada del mismo; sin embargo, es necesario aplicar las pruebas correspondientes que soporten una decisión objetiva. De tal forma que aplicaremos procedimientos para evaluar las variables en términos de cuáles deberían mantenerse en el modelo o deberían excluirse, esto lo logramos con la función “step()” en la que se pueden seleccionar los procedimientos “backward” y “forward” al anexar el argumento “direction=”. Para ejemplificar el uso de los procedimientos, utilizaremos la información resultante del modelo de regresión realizado anteriormente y guardado en la variable “MultiReg”. A continuación se aplicará la función para el procedimiento “backward”, el resultado lo guardaremos en otra variable a la que llamaremos “ModeloBackward”:

```
> ModeloBackward <- step(MultiReg, direction="backward")
Start:  AIC=19.83
Coher ~ Temp + HR + DAP
```

	Df	Sum of Sq	RSS	AIC
<none>			32.643	19.830
- DAP	1	14.797	47.440	21.569
- HR	1	39.065	71.707	25.700
- Temp	1	58.937	91.580	28.146

Aplicaciones de Estadística Básica

Aplicando el procedimiento “Backward”, el programa dio un paso y no se excluyó ninguna de las variables del modelo, de lo contrario la variable excluida no hubiesen aparecido en la tabla de resultado. El modelo sin la extracción de alguna variable (<none>), o sea con las tres variables, tiene el menor valor (19.83) de AIC (una medida de la calidad del modelo). Seguidamente aplicaremos el procedimiento “forward”, para el cual debemos separar la variable respuesta (dependiente) y predictores, para luego unirlos en el modelo. A continuación realizaremos un modelo solo para la variable respuesta y lo guardamos en la variable “Cobertura”:

```
> Cobertura <-lm(MultiVar$Cober ~ 1)
```

Luego crearemos el modelo completo anexando las variables predictoras mediante el argumento “scope=”, aplicamos el procedimiento “forward” y lo guardamos en la variable llamada “ModeloForward”:

```
> ModeloForward <-step(Cobertura, direction="forward", scope=(~  
MultiVar$Temp + MultiVar$HR + MultiVar$DAP))  
Start:  AIC=64.96  
MultiVar$Cober ~ 1
```

	Df	Sum of Sq	RSS	AIC
+ MultiVar\$Temp	1	5348.8	75.3	24.193
+ MultiVar\$HR	1	5325.2	98.9	26.918
+ MultiVar\$DAP	1	5233.0	191.1	33.502
<none>			5424.1	64.960

```
Step:  AIC=24.19  
MultiVar$Cober ~ MultiVar$Temp
```

	Df	Sum of Sq	RSS	AIC
+ MultiVar\$HR	1	27.8880	47.440	21.569
<none>		75.328	24.193	
+ MultiVar\$DAP	1	3.6202	71.707	25.700

```
Step:  AIC=21.57  
MultiVar$Cober ~ MultiVar$Temp + MultiVar$HR
```

	Df	Sum of Sq	RSS	AIC
+ MultiVar\$DAP	1	14.797	32.643	19.830
<none>		47.440	21.569	

```
Step:  AIC=19.83  
MultiVar$Cober ~ MultiVar$Temp + MultiVar$HR + MultiVar$DAP
```

Al igual que en procedimiento anterior, el programa no encontró variables para excluir del modelo, por lo que se concluye que las tres variables predictoras seguirán estando dentro del modelo.

Podemos evaluar los residuos del modelo con el uso de la función “plot()” y el modelo de regresión múltiple que fue guardado en la variable “MultiReg”:

```
> plot(MultiReg, which=1, add.smooth=FALSE, caption=NA, sub.caption=NA)
```

El gráfico de residuales no muestra signos de problemas considerables en el modelo (Figura 93), con la excepción de algunos valores extremos que pueden ser potencialmente atípicos (observaciones 4, 6 y 7).

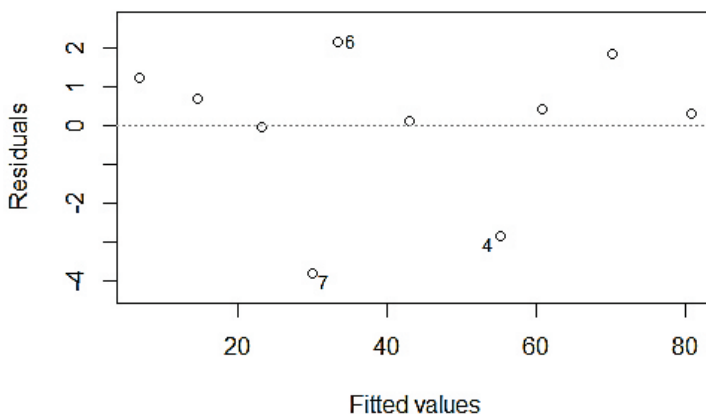


Figura 93. Gráfico de residuales para el modelo de regresión lineal múltiple.

De tal forma que la predicción de la cobertura de *Anomadon attenuatus* se haría con la ecuación:

$$\text{Cobertura \%} = 63.16 - 3.86(\text{Temperatura } ^\circ\text{C}) + 0.86(\text{Humedad Relativa \%}) - 0.68 (\text{DAP cm})$$

Por ejemplo, en un lugar donde se registre una temperatura 18 °C, una humedad relativa de 87% y un árbol de 80 cm de diámetro se esperaría encontrar una cobertura del musgo de:

$$\text{Cobertura} = 63.16 - 3.86(18) + 0.86(87) - 0.68(80)$$

$$\text{Cobertura} = 63.16 - 69.48 + 74.82 - 54.4$$

$$\text{Cobertura} = 14.2 \%$$

Opciones gráficas

Las opciones gráficas son muy importantes para representar los datos y los análisis realizados. R es un programa que “se luce” en la creación de gráficos de tipo científico. Las características de los gráficos se personalizan como en Microsoft Excel, pero a diferencia de este la personalización puede que sea un tanto más complicada para los usuarios novicios, pues se deben utilizar diferentes tipos de comandos para ejecutar cada cambio.

Es objeto de este escrito mostrar las opciones gráficas que se logran con el uso de las funciones básicas de R, no es objeto el uso de paquetes específicos. Queda en manos del lector aprender el uso de dichos paquetes, entre los que se recomiendan: MASS, gplots, plotrix, sciplot, graphics, lattice, ggplot2, entre otros.

Gráficos básicos

En este acápite se estará mostrando cómo crear gráficos versátiles y estéticos para publicaciones científicas y presentación de datos en documentos formales. Dentro de los gráficos básicos abordaremos a los gráficos de barra (una vía y dos vías); gráficos de pastel, línea y puntos; y los gráficos de barras con barras de error. Para ejemplificar el uso de las opciones gráficas se hará uso de las mismas tablas de datos que hemos venido utilizando hasta entonces, según correspondan con el tipo de gráfico.

Gráfico de barras de una vía

Es uno de los gráficos más generales y se utiliza para presentar valores numéricos en función de variables categóricas. El gráfico se obtiene con la función “`barplot()`” y el argumento son los datos a graficar. Siendo un gráfico muy común, la función tiene una vasta cantidad de argumentos, de los cuales algunos demostraremos en este escrito, más información se puede obtener ejecutando el argumento “`?barplot`” en la consola del programa. Dicha información está en inglés, si el idioma es limitante se deberá recurrir a otros recursos informativos en español. Para ejemplificar, utilizaremos los datos de humedad de suelo de 10 observaciones en tres sitios distintos, a continuación se importamos los datos a R y los guardamos en la variable que llamaremos “HS”:

```
> HS <- read.csv(file.choose())
> head(HS)
  No.Observ Sitio1 Sitio2 Sitio3
1          1    85.4    81.2    84.7
2          2    91.2    72.1    90.5
3          3    93.4    82.3    92.7
```

4	4	84.3	83.4	83.6
5	5	86.5	90.1	85.8
6	6	98.2	81.3	97.5
7	7	94.3	76.2	93.6
8	8	77.2	71.2	76.5
9	9	96.1	81.2	95.4
10	10	83.2	82.3	82.5

Primero calcularemos las medias de la humedad del suelo para cada uno de los sitios, esto se logra con la función “sapply()” y el argumento “FUN=mean”, el resultado lo guardaremos en la variable “MediasHS”.

```
> MediasHS <-sapply(HS[,2:4], FUN=mean)
> MediasHS
Sitio1 Sitio2 Sitio3
 88.98  80.13  88.28
```

El gráfico también se puede crear con un vector de los tres valores de media y la función “barplot()”, asignando los nombres con el argumento “names.arg=”. Seguidamente crearemos el gráfico de barra con la función “barplot()” (Figura 95 A):

```
> barplot(MediasHS)
```

Hemos creado un gráfico sencillo, el cual se puede personalizar con el uso de varios argumentos. Primeramente añadiremos los títulos en los ejes y el título principal del gráfico, utilizando los argumentos “xlab=” (título al eje X), “ylab=” (título al eje Y) y “main=” (título principal del gráfico) (Figura 95 B):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)",
main="Humedad del Suelo")
```

- MediasHS = Contiene las medias a ser utilizadas en el gráfico.
- xlab= Asigna el título “Sitios” al eje X.
- ylab= Asigna el título “Promedio HS(%)” al eje Y.
- main= Asigna el título principal al gráfico.

Recordemos que cada argumento se separa por una coma y que los nombres a asignar se escriben entre comillas. Seguidamente vamos a utilizar cuatro argumentos para controlar el tamaño de los elementos en el gráfico:

Aplicaciones de Estadística Básica

- `cex.axis=` Controla el tamaño de los nombres y números del eje Y.
- `cex.names=` Controla el tamaño nombres y números del eje X.
- `cex.lab=` Controla el tamaño de los títulos de los ejes.
- `cex.main=` Controla el tamaño del título principal del gráfico.

Todos los argumento por defecto presentan en el gráfico un tamaño equivalente a 1 (o 100%), de tal forma que si cambiamos a 1.5 el tamaño de los elementos incrementará un 50%, o sea será de tamaño 150%, y si cambiamos a 2 estos incrementarán un 100% y dicho elemento tendrá un tamaño de 200%. Para ejemplificar primeramente vamos a incrementar el tamaño del título del gráfico utilizando el argumento "`cex.main=`" y estableciendo su tamaño a 1.5, notar en la figura 95 C la diferencia:

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)",  
main="Humedad del Suelo", cex.main=1.5)
```

Seguidamente haremos el cambio a todas las partes del gráfico, utilizando los restantes argumentos, para los cuales vamos a establecerlos en 1.2 el contenido de los ejes y en 1.5 los títulos de los ejes:

```
> barplot (MediasHS, xlab="Sitios", y lab="Promedio HS(%)", main=  
"Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.names=1.2,  
cex.lab=1.5)
```

Observamos en la figura 95 D el incremento en los elementos dentro del gráfico, sin embargo ha surgido un problema con el título del eje Y del gráfico. Al parecer el tamaño del texto con el nombre del eje no encaja con el tamaño de los márgenes del gráfico, de tal forma que el título "Promedio HS(%)" se observa recortado. Esto lo podemos arreglar con la función "`par()`" y el argumento "`mar=`" con los cuales se controlan los márgenes. El argumento "`mar=`" que representa a los márgenes, esta denotado por un vector de cuatro número, cada número es el valor del margen comenzando desde el margen de la parte de abajo del gráfico y continuando según las manecillas del reloj (abajo, izquierda, arriba, derecha), los valores por defecto son 5, 4, 4, 2 (Figura 94).

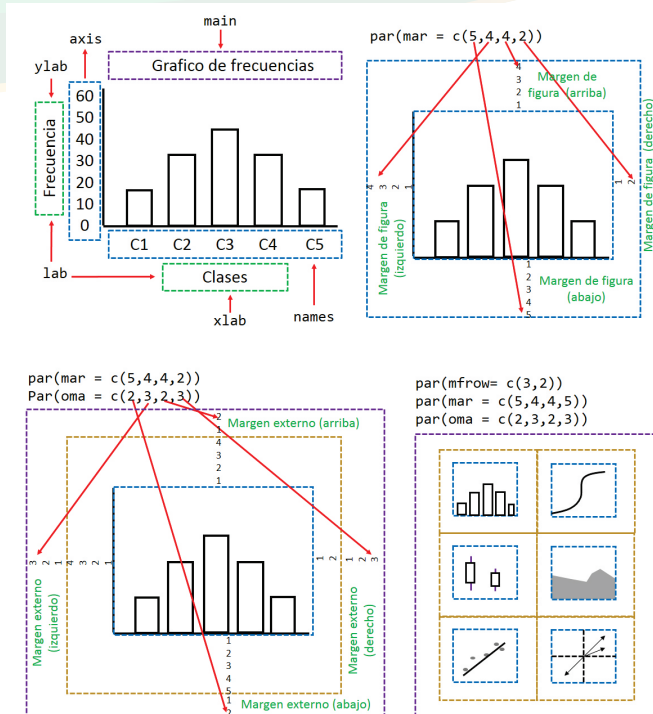


Figura 94. Arriba: a la izquierda se ilustran los nombres que toman las partes del gráfico y que son incluido en la función “`barplot()`”; a la derecha se ilustra la forma en que controlan los tamaños de los márgenes de los gráficos mediante la función “`par()`” y el argumento “`mar=`”. Abajo: a la izquierda, forma en que se añaden los márgenes externos o plantilla de gráficos mediante la función “`par()`” y el argumento “`oma=`”; a la derecha se muestra el uso del argumento “`mfrow=`” para añadir un panel de varios gráficos.

En este escrito, llamaremos “títulos de los ejes X y Y” a lo que R designa el argumento “`lab`” (`xlab`, `ylab`) abreviación de la palabra en inglés “`label`” o “etiqueta” en español. En la figura 94 arriba, el título del eje X sería “Clases” y el título del eje Y sería “Frecuencia”. A lo largo de la escala de cada eje se encuentran los “nombres y números de los ejes”, para los cuales R designa como “`names`” a los nombres o números en la escala del eje X y “`axis`” a los nombres o números en la escala del eje Y. Podemos obtener más información sobre los elementos que conforman los gráficos de barra escribiendo “`?barplot`” en la consola de R.

Los valores por defecto de todos los argumentos de la función “`par()`” los podemos conocer al escribir `par()` en la consola; en caso de querer conocer los valores por defecto de alguno de los argumentos en específico, ejemplo “`mar=`”, se ejecuta el comando “`par(“mar”)`”.

Aplicaciones de Estadística Básica

Siguiendo con el ejemplo anterior, el objetivo será hacer más grande el margen derecho para que aparezca completo el título del eje Y, para ellos cambiamos el valor del margen izquierdo del gráfico escribiendo 5 en lugar de 4, en el segundo número del vector (Figura 95 E):

```
> par(mar=c(5,5,4,2))
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)",
main="Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.
names=1.2, cex.lab=1.5)
```

Además del margen que circunda las figuras, también hay un margen de la plantilla (margen externo) que circunda los márgenes de los gráficos y estos se controlan con la función "par()" y el argumento "oma=" (Figura 94) y su esquivamente en pulgadas "omi=". Estos últimos dos argumentos tienen valor de cero por defecto, solo aparece si se hace su llamado con el comando "par(oma=())" o "par(omi=())".

Los márgenes externos son especialmente útiles cuando se tienen gráficos de paneles, o sea múltiples figuras en una sola plantilla, esto se realiza con el argumento "mfrow="; sin embargo, el uso de este último argumento será ampliamente contemplado en el tema: "Gráficos múltiples".

Seguidamente ejemplificaremos la edición de la fuente de las letras en el gráfico, para la cual se incluirán los argumentos:

- font.axis= Cambia la fuente de las letras y números en el contenido de los ejes.
- font.lab= Cambia la fuente de las letras que forman parte de los títulos de los ejes.
- font.main= Cambia la fuente de las letras que forman parte del título principal del gráfico.

Cada opción requiere la selección de una característica de las letras denotadas por números: 1 = por defecto, 2 = negrita, 3 = cursiva (itálica) y 4 = negrita y cursiva. Para ejemplificar, presentaremos en cursiva el título principal del gráfico, los contenidos de los ejes en negrita y los títulos de los ejes en cursiva y negritas. Adicionalmente demostraremos la opción "\n" la cual se utiliza cuando se tienen títulos largos y se deben colocar en dos filas, para ello aumentaremos las palabras del título principal quedando como "Humedad del Suelo en los Sitios Muestreados" y lo segmentaremos entre las palabras "en" y "los", o sea que se escribirá como "Humedad del Suelo en \n los Sitios Muestreados" (Figura 95 F):

```
> par(mar=c(5,5,4,2))
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)",
main="Humedad del Suelo en \n los Sitios Muestreados", cex.
main=1.5, cex.axis=1.2, cex.names=1.2, cex.lab=1.5, font.
axis=2, font.lab=4, font.main=3)
```

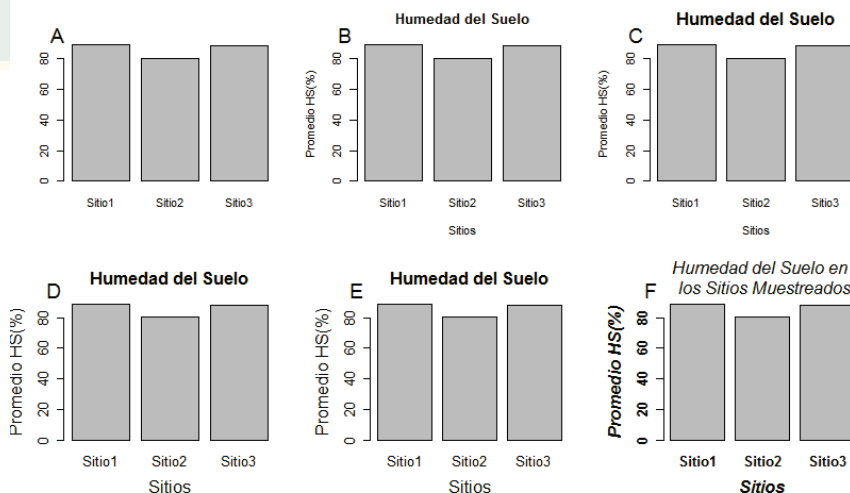


Figura 95. Panel con los gráficos de barras resultantes de la personalización con el uso de varias funciones y argumentos. A. Gráfico por defecto; B. Gráfico al que se le han anexado los títulos de los ejes y el título principal; C. Título principal agrandado; D. Títulos y contenidos de los ejes agrandados; E. Margen izquierdo del gráfico incrementado; F. Cambio de la fuente de los componentes del gráfico.

Continuando con la edición del gráfico de barra ahora controlaremos la dirección del contenido de los ejes con el argumento "las=" el cual tiene las opciones: 0 = paralelo al axis (por defecto); 1 = siempre horizontal, 2 = siempre perpendicular al axis, 3 = siempre vertical, para ejemplificar cambiaremos los números del eje Y de su posición actual (Figura 95 F) a "siempre horizontal" (1) (Figura 96 A):

```
> par(mar=c(5,5,4,2))
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)",
main="Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.
names=1.2, cex.lab=1.5, font.axis=2, font.lab=4, font.main=3,
las=1)
```

Luego pondremos los nombres del eje X en posición vertical (2) (Figura 96 B):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)",
main="Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.
names=1.2, cex.lab=1.5, font.axis=2, font.lab=4, font.main=3,
las=2)
```

Aplicaciones de Estadística Básica

Finalmente probaremos establecer la escala del eje Y y los nombres del eje X en posición vertical (3) (Figura 96 C):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)",  
main="Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.  
names=1.2, cex.lab=1.5, font.axis=2, font.lab=4, font.main=3,  
las=3)
```

Notamos que cuando los nombres del eje X se colocan en posición vertical, el gráfico no mueve el título del eje (no se autoajusta), de tal forma que tendríamos que hacer ese cambio manualmente mediante algún comando; sin embargo, para que el gráfico sea estético se prefiere mantener dichos nombres de forma horizontal a como aparece por defecto, así es que retornamos al argumento "las=1" y adicionaremos otro argumento "lwd=" para hacer un poco más grueso la escala del eje Y (lwd=2) (Figura 96 D):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)", main=  
"Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.names=1.2,  
cex.lab=1.5, font.axis=2, font.lab=4, font.main=3, las=1, lwd=2)
```

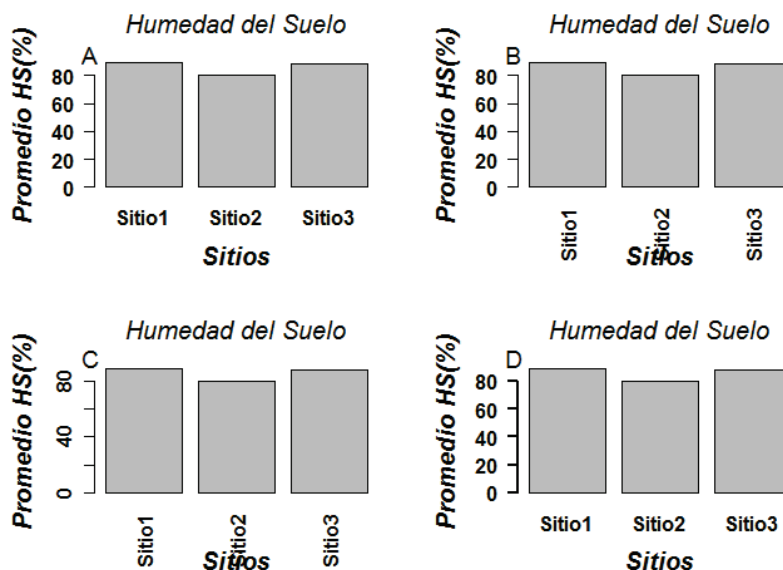


Figura 96. Panel con los gráficos de barras resultantes de personalizar la posición de los contenidos de los ejes X y Y. A. Números del eje Y en posición horizontal; B. Nombres del eje X en posición vertical; C. Números del eje Y en posición vertical y nombres del eje X en posición vertical; D. Gráfico con las condiciones como en A, pero con la escala del eje Y más gruesa.

Otras personalizaciones básicas que abordaremos en este escrito serán el controlar el grosor de las barras, de los límites y de los espacios entre barras, además de la visualización de las barras de forma horizontal. Para ejemplificar estas personalizaciones utilizaremos los mismos datos que se usaron anteriormente, correspondiente a diez observaciones de humedad relativa del suelo en tres sitios diferentes, la cual se ha guardado en la variable “HS” y cuyas medias fueron calculadas y guardadas en la variable “MediasHS”. Con las medias elaboramos el primer gráfico de barra con personalización por defecto y agregamos los títulos de los ejes X y Y (Figura 97 A):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)" )
```

Para controlar el grosor de las barras incluimos el argumento “width=” dentro del comando anterior. En la figura 97 B se observa claramente la diferencia de los grosores de las tres barras utilizando dicho argumento:

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)" ,  
width=c(1,2,5))
```

También podemos personalizar el grosor de las barras al controlar los límites del eje X, con el uso del argumento “xlim=” (Figura 97 C):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)" ,  
xlim=c(0,6))
```

Adicionalmente podemos separar las columnas utilizando el argumento “space=” y definiendo la cantidad de espacio de forma personalizada, para el ejemplo se estableció en 1 (Figura 97 D):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)" ,  
xlim=c(0,6), space=1)
```

Para tornar las barras de forma horizontal, quitaremos los argumentos “xlim=” y “space=” a fin de no tornar muy complejo el comando, seguidamente utilizamos el argumento “horiz=TRUE”. Adicionalmente pondremos los nombres del eje Y en posición horizontal con el argumento “las=” (Figura 97 E):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)" ,  
horiz=TRUE, las=1)
```

Notamos que se presenta un problema, pues el título del nuevo eje Y se superpone con los nombres del eje (Sitio 1, Sitio 2 y Sitio 3), para brindarle solución seguiremos los siguientes pasos:

Aplicaciones de Estadística Básica

1. Aumentamos el margen izquierdo: Esto lo logramos cambiando los márgenes con la función "par()" y el argumento "mar=" y aumentamos un valor al segundo número pasándolo de 4.1 a 5.1. Recordando que cada número separado por coma es el valor del margen, comenzando con el margen de abajo, siguiendo con el margen izquierdo y continuando con el margen superior y margen derecho (abajo, izquierdo, arriba, derecho).
2. Reubicamos el título del eje Y al eje X: Para ello simplemente utilizamos el argumento "xlab" y asignamos el título correspondiente.
3. Suprimimos el título del eje Y: Escribiendo "NA" (equivalente a ninguno) en el argumento "ylab=".
4. Colocamos el nuevo título de Y como una anotación: Dado a que el gráfico no tiene título del eje Y, este lo asignamos de una forma flexible de personalización utilizando la función "title()" y agregando dos argumentos, el primero define el título del eje Y y el segundo le indica al programa el número de línea en que se establecerá el título del eje Y. Para el caso de este ejemplo se estableció en la línea 4 quedando a una línea de distancia de las etiquetas del mismo eje.

Los resultados finales se observan en la figura 97 F y las tres líneas de comando se muestran a continuación:

```
> par(mar=c(5.1,5.1,4.1,2.1))  
> barplot(MediasHS, xlab="Promedio HS(%)", ylab=NA, horiz=TRUE, las=1)  
> title(ylab="Sitios", line=4)
```

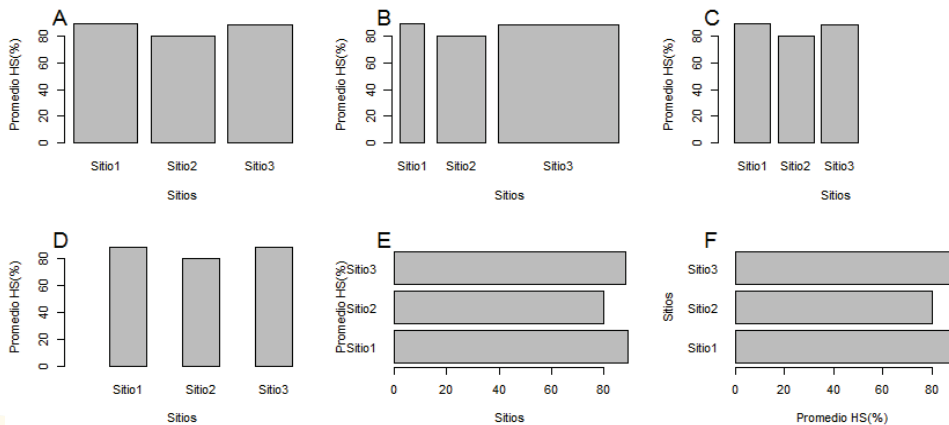















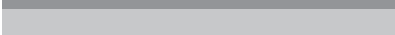









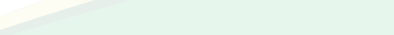
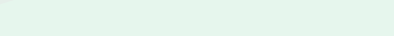
Figura 97. Personalización extra de un gráfico de barra. A. Gráfico con opciones por defecto y títulos de ejes; B. Personalización del grosor de las barras; C. Cambio del ancho de las barras; D. Control de la distancia entre las barras; E. Cambio a barras horizontales; F. Modificación del título y de los nombres del eje Y para remediar superposición.

En las publicaciones científicas se prefiere generalmente gráficos sencillos, de colores blanco y negro o tonos de grises, los gráficos muy coloridos no se suelen usar en publicaciones científicas (en dependencia de la revista o editorial) y por lo general se utilizan solamente para ilustrar presentaciones orales, de tal forma que en este documento abordaremos algunas funciones para asignar color a los gráficos.

A como hemos notado, para personalizar los gráficos es preciso anexar argumentos a la función para crear el gráfico, de igual forma se asignan los colores. Asignaremos los colores como argumentos referidos al nombre de cada color escrito en inglés (Cuadro 18), adicionalmente también se suele utilizar una codificación en números de tal forma que un número está asociado a un color. Si deseamos ver la lista de colores se puede escribir en la consola la función “colors()” y se presiona Enter; si quisiéramos ver todos los colores disponibles en R y sus códigos, podemos descargarlos de <http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>.

Cuadro 18. Algunos colores de uso común aplicados en R con el argumento “col=”

Color	Código	Nombre
	black	Negro
	blue	Azul
	chartreuse	Verde claro
	chartreuse4	Verde
	darkblue	Azul oscuro
	darkgreen	Verde oscuro
	darkred	Rojo oscuro
	darkviolet	Violeta oscuro
	deeppink	Rosado intenso
	deepskyblue	Celeste intenso

Color	Código	Nombre
	gold	Color oro
	gray	Gris
	gray25	Gris 25%
	gray50	Gris 50%
	gray75	Gris 75%
	green	Verde
	midnightblue	Azul de media noche
	navy	Azul de la Armada
	orange	Naranja
	purple	Púrpura
	red	Rojo
	saddlebrown	Marrón
	violet	Violeta
	white	Blanco
	yellow	Amarillo

Aplicaciones de Estadística Básica

Podemos asignar color, tanto al gráfico como a sus componentes, por ejemplo contenidos y título de los ejes:

- `col=` Da color a las barras.
- `col.axis=` Da color al contenido de los ejes.
- `col.lab=` Asigna color a los títulos de los ejes.
- `col.main=` Asigna color al título principal.

A continuación asignaremos colores al gráfico que elaboramos anteriormente, utilizando los códigos del cuadro 18. A modo de ejemplo asignaremos color azul (`blue`) a las barras del gráfico, gris (`gray50`) a los contenidos de los ejes, verde oscuro (`darkgreen`) al título principal y rojo oscuro (`darkred`) a los títulos de los ejes. Dichos códigos se especifican entre comillas (Figura 98 A):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)", main="Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.names=1.2, cex.lab=1.5, font.axis=2, font.lab=4, font.main=3, las=1, lwd=2, col="blue", col.axis="gray50", col.lab="darkgreen", col.main="darkred")
```

También podemos controlar el color de los bordes utilizando el argumento `"border="` y cambiar los nombres de las etiquetas en X con el argumento `"names="` y un vector con la lista de nombres (Figura 98 B):

```
> barplot(MediasHS, xlab="Sitios", ylab="Promedio HS(%)", main="Humedad del Suelo", cex.main=1.5, cex.axis=1.2, cex.names=1.2, cex.lab=1.5, font.axis=2, font.lab=4, font.main=3, las=1, lwd=2, col="blue", col.axis="gray50", col.lab="darkgreen", col.main="darkred", border="red", names=c("Lugar 1", "Lugar 2", "Lugar 3"))
```

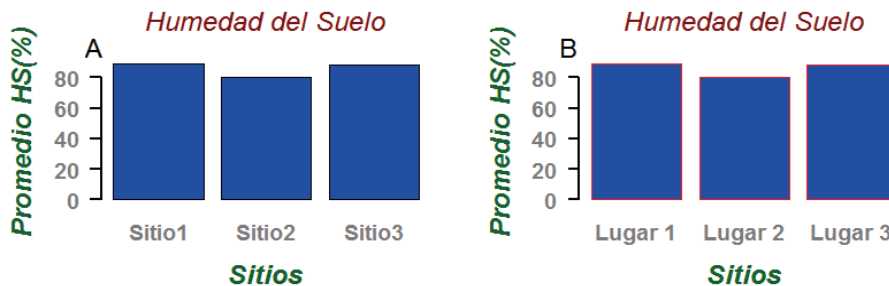


Figura 98. Gráficos de barra con colores. A. Resultante de la asignación de color con la serie de argumentos `"col="`, `"col.axis="`, `"col.lab="`, `"col.main="` y `"border="`; B. Igual que A, pero con diferentes nombres en el eje X.

Finalmente estableceremos los valores de las medias en cada barra a modo de capa, para esto guardaremos el último gráfico que hemos creado (Figura 98 B) en una variable llamada "Grafico". Luego utilizaremos la función "text()" para insertar los valores. En la función "text()" se asignan varios argumentos, los principales son cuatro, el primero es el nombre de la variable donde se guardó el gráfico (Grafico); el segundo argumento es un número que define la posición que van a estar los valores a asignar con respecto a los números de la escala del eje Y, en nuestro caso los asignaremos a la altura del número 17 (número de referencia); el tercer argumento es "pos=" con el que definimos la posición con respecto al número de referencia, las opciones son 1 que representa abajo, 2 a la izquierda, 3 arriba y 4 a la derecha, para este ejemplo lo pondremos abajo o sea "pos=1"; el cuarto argumento es "paste()" con los que se pegan los números en cada barra, los cuales fueron previamente guardados en la variable "MediasHS".

Otros argumentos que se incluirán en este ejemplo son: "col=" para asignar el color a los números y "font=" para cambiar la fuente (Figura 99). Adicionalmente también se puede incluir el argumento "cex=" para cambiar el tamaño de los números.

```
> Grafico <-barplot(MediasHS, xlab="Sitios", ylab="Promedio  
HS(%)", main="Humedad del Suelo", cex.main=1.5, cex.axis=1.2,  
cex.names=1.2, cex.lab=1.5, font.axis=2, font.lab=4, font.  
main=3, las=1, lwd=2, col="blue", col.axis="gray50", col.lab=  
"darkgreen", col.main="darkred", border="Red", names=c("Lugar  
1", "Lugar 2", "Lugar 3"))  
> text(Grafico, 17, pos=1, paste(MediasHS), col="white", font=2)
```

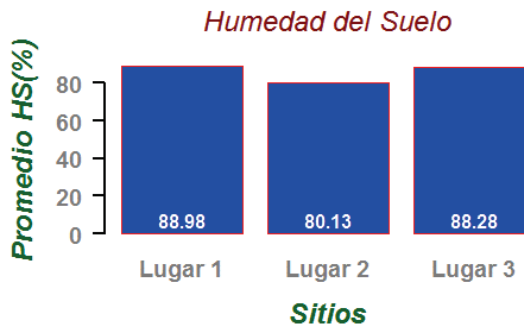


Figura 99. Gráfico de barra con los valores de las medias asignados respectivamente.

En todos los gráficos podemos también incluir anotaciones a modo de capas, esto es importante para añadir información extra a los gráficos, que sean útiles para explicar situaciones especiales. Esto lo logramos con las funciones "text()" y "mtext()". Para utilizar la función "text()" se tienen que especificar las coordenadas X y Y dentro del gráfico, la coordenada X se obtiene de algún valor tomado del eje X y la coordenada Y se obtiene de algún valor tomado del eje Y (Figura 100 B). Ejemplificaremos el uso de la función

Aplicaciones de Estadística Básica

insertando un par de textos en el gráfico anteriormente creado (Figura 99) y guardado en la variable “Grafico”, estos textos son N1 y N2 (N= Nota), ambas en diferentes posiciones, N1 en la posición $x=1$ y $y=80$ y N2 un poco más a la derecha y más abajo de posición de N1, específicamente en $x=0.7$, $y=65$. Adicionaremos el argumento “col=” y seleccionaremos el color amarillo (yellow) para los textos. La adición de los dos textos se realiza en la tercera y cuarta línea del siguiente comando:

```
> Grafico
> text(Grafico, 17, pos=1, paste(MediasHS), col="white", font=2)
> text(x=1, y=80, label="N1", col="yellow")
> text(x=0.7, y=65, label="N2", col="yellow")
```

En la figura 100 A se muestran los dos textos (N1 y N2) añadidos en color amarillo; adicionalmente en la figura 100 B se ilustran las coordenadas utilizadas para adicionar cada texto. Como el eje X tiene elementos categóricos (nombres) la posición $x=1$ se ubica en el margen de la barra del “Sitio 1”, si se desea que la etiqueta aparezca sobre la segunda o tercera barra (Sitio 2 o Sitio 3), se deberán establecer las coordenadas a $x=2$ o $x=3$. Es recomendable que el usuario pruebe diferentes posiciones jugando con las coordenadas hasta que el texto esté ubicado en el lugar deseado.

Con la función “mtext()” también se agregan textos, pero fuera de los márgenes del gráfico. En lugar de definir coordenadas, se define el lado del gráfico en donde se pretende desplegar el texto. Los lados son abajo= 1; lado izquierdo= 2; arriba= 3 y lado derecho= 4. Para demostrar su uso, insertaremos unas notas (Nota A, Nota B, Nota C y Nota D) en cada uno de los lados del gráfico respectivamente; adicionalmente le asignaremos color rojo (red) al texto (Figura 100 C):

```
> Grafico
> text(Grafico, 17, pos=1, paste(MediasHS), col="white", font=2)
> mtext(text="Nota A", side=1, col="red")
> mtext(text="Nota B", side=2, col="red")
> mtext(text="Nota C", side=3, col="red")
> mtext(text="Nota D", side=4, col="red")
```

Evidentemente las notas se sobrepondrán sobre otros elementos del gráfico, por lo tanto el usuario debe de juzgar en qué posición es más favorable establecer los textos. La función “mtext()” también nos ofrece un argumento para establecer el texto no en el centro de cada lado sino en los extremos de cada lado, este argumento es “adj=” (“adjustment” o ajuste en español) con dos opciones: 0 establece el texto en el extremo izquierdo (cuando se establece en los lados de abajo y arriba del gráfico) y en el extremo de abajo (cuando se establece en los lados de izquierdo y derecho del gráfico); 1 establece

el texto en el extremo opuesto al que se establece con la opción 0. Para ejemplificar, añadiremos dos textos en el lado del gráfico (abajo), uno dirá “Nota A_0” e irá en extremo derecho de ese lado y el otro dirá “Nota A_1” e irá en el extremo derecho del lado 1 (Figura 100 D), le asignaremos color rojo (red) a los textos:

```
> Grafico
> text(Grafico, 17, pos=1, paste(MediasHS), col="white", font=2)
> mtext(text="Nota A_0", side=1, adj=0, col="red")
> mtext(text="Nota A_1", side=1, adj=1, col="red")
```

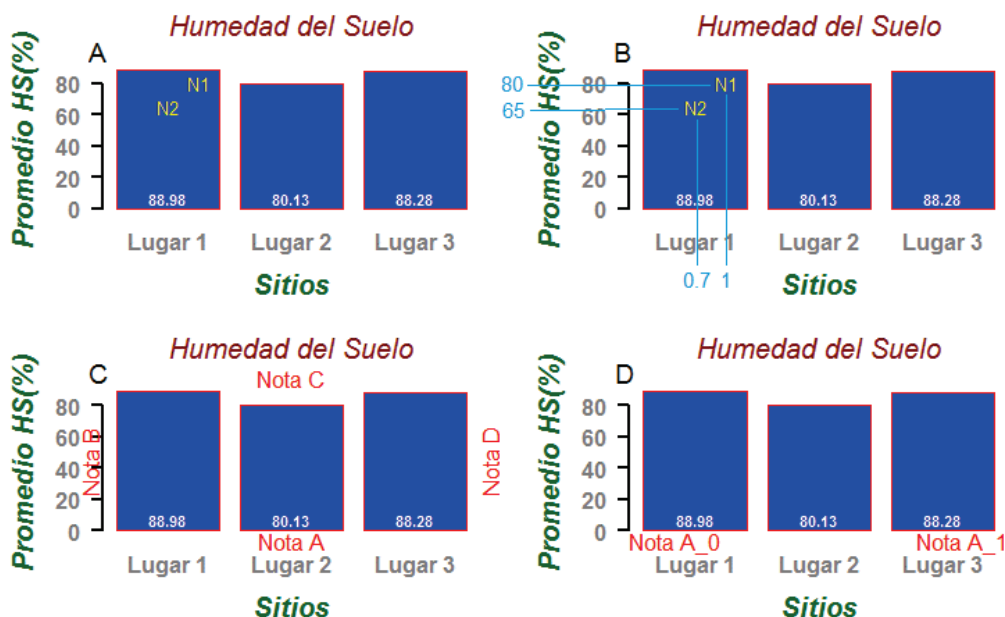


Figura 100. Demostración de la inserción de texto dentro de los gráficos. A. Inserción de textos llamados N1 y N2; B. Ilustración de las coordenadas asignadas a dichos textos; C. Inserción de texto fuera de los márgenes del gráfico; D. Demostración de la posición izquierda y derecha de los textos en el margen 1 (abajo).

Gráfico de barras de dos vías

En algunos casos es necesario crear gráficos de barras de doble vía (o barras agrupadas), en especial cuando tenemos dos variables categóricas con varios niveles. Todas las opciones de edición que se estudiaron anteriormente aplican para estos gráficos. Para ejemplificar haremos uso de unos datos de pH de suelo tomados en diferentes puntos, dentro de dos ecosistemas (Bosque y Agrícola) presentes en las tres partes de

Aplicaciones de Estadística Básica

una microcuenca (Alta, Media, Baja) (toposecuencia). Importaremos los datos a R y los guardaremos en una variable llamada “pH”:

```
> pH <-read.csv(file.choose())
```

```
> pH
```

	Ecosist	Alta	Media	Baja
1	Bosque	4.5	4.3	3.1
2	Bosque	5.3	3.1	4.0
3	Bosque	4.3	2.3	2.3
4	Bosque	4.8	3.4	2.4
5	Bosque	4.7	2.1	3.1
6	Agrícola	3.7	3.5	2.3
7	Agrícola	4.5	2.3	3.2
8	Agrícola	3.5	1.5	1.5
9	Agrícola	4.0	2.6	1.6
10	Agrícola	3.9	1.3	2.3

Seguidamente es necesario calcular las medias por ecosistema y por cada parte de la microcuenca. Para ello extraeremos las medias de los datos en formato vectorial y luego la transformaremos a un formato de tabular. La extracción la realizaremos con la función “sapply()” y como argumento el número de filas que corresponden al ecosistema Bosque (1:5) y el rango de columnas de la variable toposecuencia (2:4), la información resultante la guardaremos en una variable a la que llamaremos “Bosque”:

```
> Bosque <-sapply(pH[1:5,2:4], FUN=mean)
```

```
> Bosque
```

Alta	Media	Baja
4.72	3.04	2.98

Aplicaremos el mismo procedimiento al segundo ecosistema (Agrícola) y guardamos el resultado en la variable “Agricola” (para facilitar la escritura en la consola obviaremos temporalmente el acento):

```
> Agricola <-sapply(pH[6:10,2:4], FUN=mean)
```

```
> Agricola
```

Alta	Media	Baja
3.92	2.24	2.18

A continuación combinamos los dos vectores mediante la función “rbind()” que combina datos por medio de las filas (opuesto a “cbind()” que combina datos por medio de las columnas), la información la guardaremos en una nueva variable llamada “Datos”:

```
> Datos <-rbind(Bosque,Agricola)
> Datos
      Alta Media Baja
Bosque  4.72   3.04 2.98
Agricola 3.92   2.24 2.18
```

Finalmente hemos creado la tabla de datos con las medias de pH para ser utilizados en la elaboración del gráfico de barras de dos vías con el uso de la función “barplot()” (Figura 101 A):

```
> barplot(Datos)
```

La función genera por defecto un gráfico de barras apiladas, que modificaremos para poner una barra al lado de la otra, dicha modificación se hace con el argumento “beside=TRUE” (Figura 101 B):

```
> barplot(Datos, beside=TRUE)
```

A continuación agregaremos los títulos de los ejes y estableceremos los números de la escala del eje Y de forma horizontal (Figura 101 C):

```
> barplot(Datos, beside=TRUE, xlab="Toposecuencia",ylab=
"Promedios de pH", las=1)
```

Adicionalmente añadiremos colores a cada categoría de la variable ecosistema, a la categoría “Bosque” le asignaremos color azul (blue) y a la categoría “Agrícola” el color verde (green), esta asignación se realizará con el argumento “col=” y especificando los colores mediante un vector estructurado con la función “c()” (Figura 101 D):

```
> barplot(Datos, beside=TRUE, xlab="Toposecuencia", ylab=
"Promedios de pH", las=1, col=c("blue","green"))
```

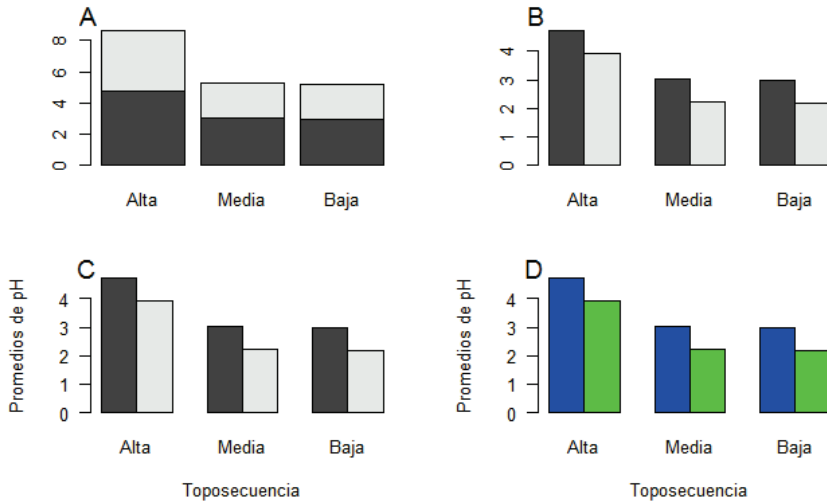


Figura 101. Gráficos mostrando las personalizaciones realizadas a un gráfico de barras con dos barras juntas.

Cuando el número de categorías es muy grande dentro de una variable categórica, es muy tedioso el escribir todos los colores mediante un vector, de tal forma que es más conveniente utilizar las paletas de colores básicas que tiene R, que se usan con la función “col=”, algunos ejemplos son:

Paleta de colores grises: `gray.colors (n, start=0, end=1, gamma=1)`

Donde,

n= Números de colores a utilizar de la paleta. Referido también al número de categorías dentro de la variable categórica a representar.

start= En qué tono de la paleta se iniciará la asignación de los colores, siendo 0 el tono completamente negro.

end= En qué tono de la paleta se finalizará la asignación de los colores, siendo 1 el tono completamente blanco. Los valores intermedio se asignan con números decimales, ejemplo: 0.75, 0.50, 0.25, etc.

En Microsoft® Excel y R

gamma= Controla la tonalidad oscura de los colores de la paleta.

Paleta de colores de arcoíris: rainbow(n, start=0, end=1, alpha=1)

Donde,

start= Color inicial de la paleta.

end= Color final de la paleta.

alpha= Transparencia alfa.

Otras paletas: heat.colors(n, alpha=1), terrain.colors(n, alpha=1), topo.colors(n, alpha=1), cm.colors(n, alpha=1).

Continuando con la personalización del gráfico de pH de suelo, vamos a insertar en el gráfico la leyenda para que nos informe sobre el significado de cada color. Hay varias formas de asignar la leyenda, la primera es simplemente utilizando el argumento "legend=TRUE" (Figura 102 A):

```
> barplot(Datos, beside=TRUE, xlab="Toposecuencia", ylab="Promedios de pH", las=1, col=c("blue", "green"), legend=TRUE)
```

La ventaja de este argumento es que colocará rápidamente la leyenda en la parte superior derecha del gráfico y presenta el nombre de las filas al lado de cada color; la desventaja es que el tamaño de la leyenda y su posición podrían no ser adecuados para la estética del gráfico.

Otra forma de asignar leyenda dentro de la función "barplot()" es el uso de la función "legend()", con esta especificamos la posición en la que queremos la leyenda dentro del gráfico; además, se puede agregar el nombre y el color del contenido de la leyenda. La posición la definimos con los argumentos: topleft= parte superior izquierda, topright= parte superior derecha, bottomleft= parte inferior izquierda, bottomright= parte inferior derecha, center= centro del gráfico.

Para ejemplificar, colocaremos la leyenda en la parte superior izquierda del gráfico (topleft), además anexaremos los nombres de los elementos de la leyenda con el argumento "legend=" y asignaremos los colores con el argumento "fill=" (Figura 102 B):

```
> barplot(Datos, beside=TRUE, xlab="Toposecuencia", ylab="Promedios de pH", las=1, col=c("blue", "green"), legend("topleft", legend=c("Bosque", "Agricola"), fill=c("blue", "green")))
```

Aplicaciones de Estadística Básica

La desventaja de esta función es que el usuario solo tiene cinco opciones para posicionar la leyenda, esto puede ser problemático si todas estas posiciones posibles se superponen con el gráfico (Figura 102 C).

Para tener más libertad en seleccionar la posición donde queremos colocar la leyenda, sustituimos el nombre de la posición en inglés (topleft, topright...), por el argumento "locator(1)", con esto el programa nos dará la oportunidad que con un clic sobre el área del gráfico se coloque la leyenda. Es necesario seleccionar previamente el área donde pueda caber la leyenda sin que se sobreponga con las figuras del gráfico, para esto último también es posible reducir el tamaño de la leyenda con el argumento "cex=" dentro de la función "legend()".

El argumento requiere que se señalice donde se tomarán los nombres de la leyenda, por lo que se usa la función "rownames()" con el cual se colocará el nombre de las filas de la tabla de datos (Bosque y Agrícola), el color se mantendrá igual con el argumento "fill=" (Figura 102 D):

```
> barplot(Datos, beside=TRUE, xlab="Toposecuencia", ylab="Promedios de pH", las=1, col=c("blue","green"), legend(locator(1), rownames(Datos), fill=c("blue","green"), cex=0.7))
```

Para mejorar la estética de la leyenda, podemos agregar un título a la misma, esto se logra con el argumento "title=" dentro de la función "legend()". Agregaremos el título "Ecosistemas" a la leyenda recién creada (Figura 102 E):

```
> barplot(Datos, beside=TRUE, xlab="Toposecuencia", ylab="Promedios de pH", las=1, col=c("blue","green"), legend(locator(1), rownames(Datos), fill=c("blue","green"), cex=0.7, title="Ecosistemas"))
```

Otro argumento importante que la función "legend()" tiene disponible para la función "barplot()" es colocar los componentes de la leyenda uno al lado de otro, en lugar de uno debajo del otro, esto se logra con el argumento "horiz=TRUE" (Figura 102 F):

```
> barplot(Datos, beside=TRUE, xlab="Toposecuencia", ylab="Promedios de pH", las=1, col=c("blue","green"), legend(locator(1), rownames(Datos), fill=c("blue","green"), cex=0.7, title="Ecosistemas", horiz=TRUE))
```

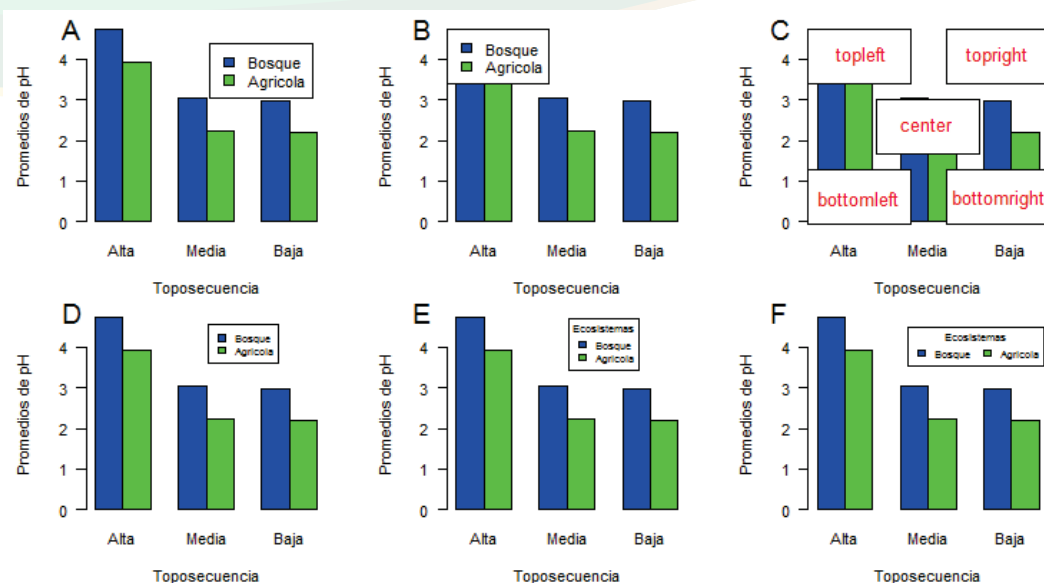



Figura 102. Ilustración de las diferentes formas de insertar y editar la leyenda de un gráfico. A. Se utiliza el argumento “`legend = TRUE`”; B. Uso de la función “`legend()`” y se especifica la posición de la leyenda en el lado superior izquierdo (`topleft`) del gráfico; C. Diferentes posiciones donde se puede colocar la leyenda; D. Uso del argumento “`locator(1)`” para seleccionar la posición de una forma libre y el argumento “`cex=`” para controlar el tamaño de la caja de la leyenda; E. Uso del argumento “`title=`” para insertar un título a la leyenda; F. El argumento “`horiz=`” permite poner los componentes de la leyenda unos bajo otros o unos al lado de otros.

Si las etiquetas tienen muchos elementos, estos se pueden arreglar en columnas con el argumento “`ncol=`” dentro de la función “`legend()`”, las opciones dependen del número de columnas en las que uno desea que se arreglen dichos elementos.

Gráfico de pastel

Los gráficos de pastel se utilizan para representar frecuencias o proporciones con respecto a valores globales. Aunque es un gráfico popular, no es muy apreciado en publicaciones científicas en especial porque no representa claramente los valores cuando el pastel está muy “seccionado” (muchas categorías dentro de la variable categórica), en cuya circunstancia se recomienda en su lugar el uso de gráficos de barra. A pesar de lo anterior, el gráfico de pastel se ha incluido dentro de este escrito como una opción más. Para hacer uso del gráfico necesitamos una lista de características con sus correspondientes frecuencias. Como ejemplo utilizaremos una lista de clases taxonómicas, a las cuales pertenecen 120 especies, las clases son mamíferos, aves, reptiles y anfibios y el

Aplicaciones de Estadística Básica

objetivo es representar en un gráfico de pastel cuantas especies hay de cada clase. Los datos que se presentarán no muestran las especies, pero sí muestran las clases a las que pertenecían cada especie. A como hacemos rutinariamente, primeramente importamos los datos a R y los guardamos en una variable a la que llamaremos “Clases” (la lista completa se muestra en el anexo 5):

```
> Clases <-read.csv(file.choose())
> head(Clases)
  Clases
1 Mamífero
2 Mamífero
3 Mamífero
4 Mamífero
5 Mamífero
6 Mamífero
> tail(Clases)
  Clases
115 Reptil
116 Reptil
117 Reptil
118 Reptil
119 Reptil
120 Reptil
> unique(Clases$Clases)
[1] Mamífero Ave      Anfibio  Reptil
Levels: Anfibio Ave Mamífero Reptil
```

La columna de datos se transforma en una tabla de frecuencia, utilizando la función “table()” y la guardamos en una nueva variable llamada “ClasesT”:

```
> ClasesT <-table(Clases$Clases)
ClasesT
Anfibio      Ave      Mamífero      Reptil
      6      88      12      14
```

Otra forma de ingresar los datos es mediante el uso de vectores, creamos un vector con los nombres de las clases (anfibios, aves, mamíferos y reptiles), otro vector con las cuentas (6, 88, 12 y 14), para combinar ambos vectores utilizamos la función “rbind()” y los guardamos en una variable, pero esta vía conlleva mucha codificación, sin embargo ambas vías llevan al mismo resultado.

Una vez generada la tabla de datos, se procede a crear el gráfico de pastel, poco a poco iremos personalizando el gráfico. Iniciaremos creando el gráfico por defecto con la función “pie()” (Figura 103 A):

```
> pie(ClasesT)
```

Luego incrementaremos el tamaño a su tamaño máximo (1) mediante el argumento “radius=” (Figura 103 B):

```
> pie(ClasesT, radius=1)
```

A continuación especificaremos el ángulo inicial para la formación del pastel con el argumento “init.angle=”, en este caso especificaremos (90°) (Figura 103 C):

```
> pie(ClasesT, radius=1, init.angle=90)
```

También podemos ordenar el gráfico de la porción de pastel más grande a la más pequeña, siguiendo las manecillas del reloj con el argumento “clockwise=TRUE” (Figura 103 D):

```
> pie(ClasesT, radius=1, init.angle=90, clockwise=TRUE)
```

Agregaros los colores mediante la función “col=” y un vector con los nombres de los colores, según la tabla de colores del cuadro 18. En este caso asignaremos el color “deeppink” (rosado intenso) a la porción de anfibio; “deepskyblue” (celeste intenso) a la porción de Ave; “gold” (color oro) a mamífero y “chartreuse” (verde claro) a Reptil (Figura 103 E):

```
> pie(ClasesT, radius=1, init.angle=90, clockwise=TRUE,  
col=c("deeppink", "deepskyblue", "gold", "chartreuse"))
```

Finalmente agregamos un título del gráfico con el argumento “main=” (Figura 103 F):

```
> pie(ClasesT, radius=1, init.angle=90, clockwise=TRUE,  
col=c("deeppink", "deepskyblue", "gold", "chartreuse"), main=  
"Clases Taxonómicas")
```

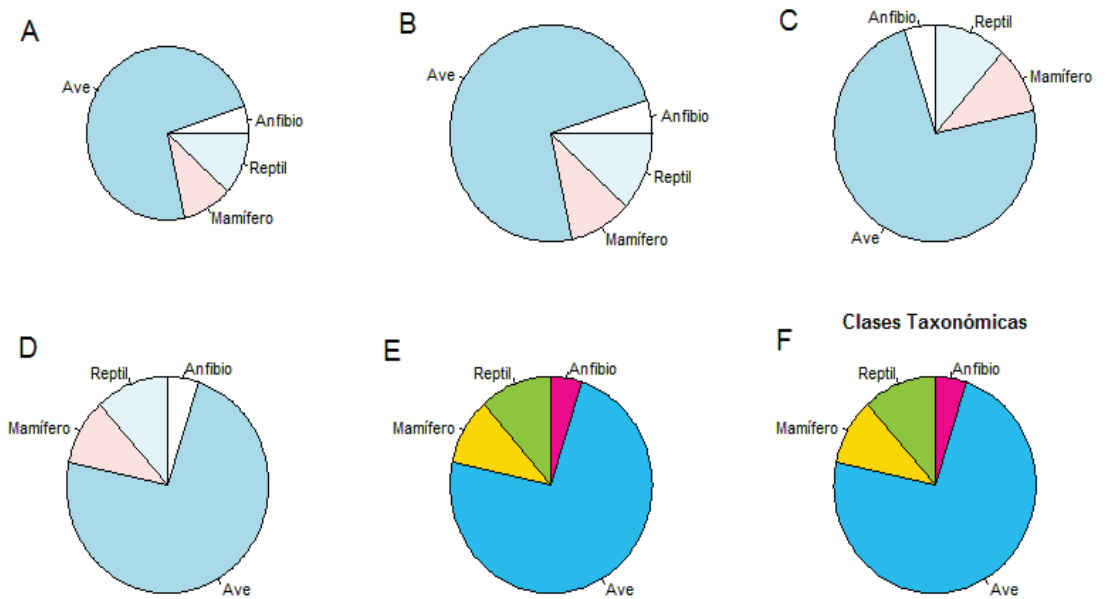


Figura 103. Ilustración de la inserción y personalización de un gráfico de pastel. A. Gráfico por defecto; B. Gráfico agrandado a su tamaño máximo; C. Pastel iniciando con la porción con ángulo aproximado a 90°; D. Porciones organizadas de mayor a menor siguiendo las manecillas del reloj; E. Asignación de colores; F. Asignación de título principal.

A continuación haremos un cambio de las etiquetas de cada porción del pastel, en lugar de presentar los nombres, presentaremos el porcentaje de cada una y agregaremos una leyenda que indique el nombre de cada una de las clases. Primeramente se presentará el gráfico básico elaborado anteriormente (Figura 104 A):

```
> pie(ClasesT, radius=1, init.angle=90, clockwise=TRUE,
col=c("deeppink", "deepskyblue", "gold", "chartreuse"))
```

Al gráfico creado con el comando anterior le asignaremos las nuevas etiquetas, para ellos convertiremos los números presentes en la tabla de la variable "Clases" en un solo vector numérico, con el uso de la función "as.vector()" y se guardarán en una nueva variable llamada "Etiquetas":

```
> Etiquetas <- as.vector(ClasesT)
> Etiquetas
[1] 6 88 12 14
```

Ahora asignaremos la nueva etiqueta al gráfico de pastel, utilizando el argumento "labels=" (Figura 104 B):

```
> pie(ClasesT, radius=1, init.angle=90, clockwise=TRUE,
col=c("deeppink", "deepskyblue", "gold", "chartreuse"),
labels=Etiquetas)
```

Notamos que ahora el gráfico presenta las frecuencias en lugar de los nombres de las clases. Con este paso podemos calcular los porcentajes, utilizando el comando "round(Etiquetas/sum(Etiquetas) * 100, 1)" en el cual el argumento "Etiquetas/sum(Etiquetas) * 100" le indica al programa que tome todos los valores de la variable "Etiquetas" (6, 88, 12 y 14) y divida cada uno entre la suma de los mismos números (120). El resultado de dicha división se multiplica por 100 (* 100) para mostrar los porcentajes. Adicionalmente, se indica al programa que los resultados se redondeen a un decimal, para ellos utilizamos el argumento "round(Etiquetas/sum(Etiquetas) * 100, 1)" y el número final, que es el 1, es el que indica el número de decimales. Toda la información se guardará en la variable "Etiquetas2".

```
> Etiquetas2 <-round(Etiquetas/sum(Etiquetas) * 100, 1)
> Etiquetas2
[1] 5.0 73.3 10.0 11.7
```

Antes de finalizar es necesario, por estética, colocar los símbolos de porcentaje (%) a cada uno de los valores. Esto lo logramos con la función "paste()" y utilizamos tres argumentos, el primero es la variable donde se encuentran almacenados los valores porcentuales (Etiquetas2), el segundo es el símbolo que se desea pegar a cada valor (%) y el tercer argumento (sep=""), le indica al programa que entre cada valor y su símbolo no haya espacios, no escribir este argumento incurre en que el programa colocará un espacio de forma automática entre los valores y los símbolos. Los resultados se guardan en otra variable ahora llamada "Etiquetas3":

```
> Etiquetas3 <-paste(Etiquetas2, "%", sep="")
> Etiquetas3
[1] "5%" "73.3%" "10%" "11.7%"
```

Finalmente añadimos los valores porcentuales, con sus símbolos (guardados en la variable Etiquetas3) al comando para crear el gráfico de pastel utilizando el argumento "labels=". Como una característica extra, también aprovechamos para incrementar el tamaño de las etiquetas, con el uso del argumento "cex=" (Figura 104 C):

```
> pie(ClasesT, radius=1, init.angle=90, clockwise=TRUE,
col=c("deeppink", "deepskyblue", "gold", "chartreuse"),
labels=Etiquetas3, cex=1.5)
```

Aplicaciones de Estadística Básica

A continuación, añadimos la leyenda con la función “legend()” más el argumento “locator(1)” a fin de tener flexibilidad para colocar la leyenda en cualquier parte del gráfico. Notemos que el comando para generar la leyenda se escribe aparte del comando para crear el gráfico (o sea en diferente línea de comando) a modo de capa. Los argumentos que siguen después de “locator(1)” dentro de la función “legend()”, asignan los nombres y los colores a la leyenda (Figura 104 D):

```
> pie(ClasesT, radius=1, init.angle=90, clockwise=TRUE,  
col=c("deeppink", "deepskyblue", "gold", "chartreuse"), labels=  
Etiquetas3, cex=1.5)  
> legend(locator(1), c("Anfibios", "Aves", "Mamíferos", "Reptiles"),  
fill=c("deeppink", "deepskyblue", "gold", "chartreuse"))
```

A lo inmediato notamos que en este gráfico el espacio para colocar la leyenda no es suficiente, para solucionar este conflicto, no controlaremos el tamaño del borde del gráfico, sino el tamaño del gráfico en sí. Ya que establecimos el tamaño del gráfico con radio igual a 1 (radius=1), reduciremos ese radio a 0.6, así quedará más espacio libre entre el pastel y el borde de la plantilla del gráfico, en el cual se puede colocar la leyenda, la cual también la reducimos a 0.7 con “cex=” (Figura 104 E):

```
> pie(ClasesT, radius=0.6, init.angle=90, clockwise=TRUE,  
col=c("deeppink", "deepskyblue", "gold", "chartreuse"), labels=  
Etiquetas3, cex=1.5)  
> legend(locator(1), c("Anfibios", "Aves", "Mamíferos", "Reptiles"),  
fill=c("deeppink", "deepskyblue", "gold", "chartreuse"), cex=0.7)
```

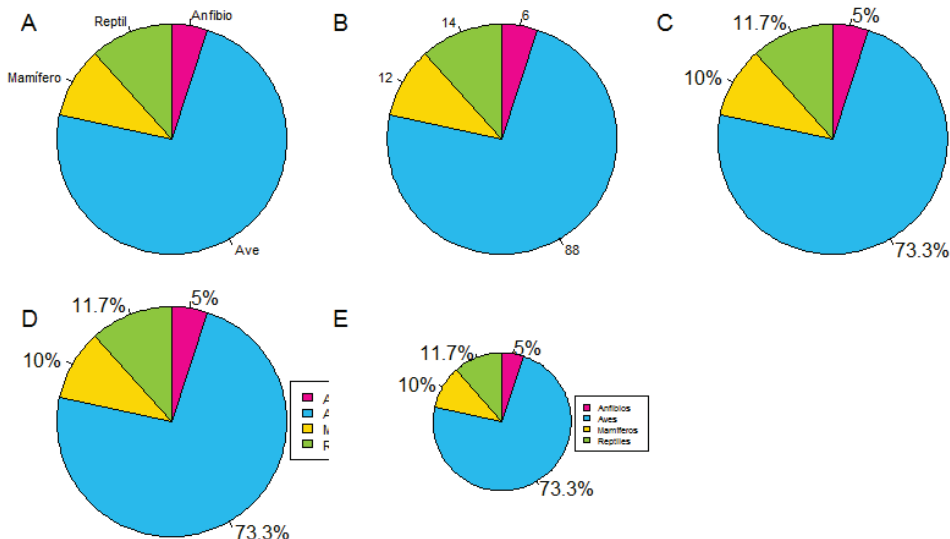


Figura 104. Ilustración del proceso para cambiar las etiquetas de cada una de las porciones del gráfico de pastel y de añadir la leyenda. A. Gráfico creado anteriormente e ilustrado como producto final en la figura 103 F; B. Gráfico con los valores numéricos de cada porción como etiquetas; C. Gráfico con los valores porcentuales de cada porción como etiquetas; D. Asignación de la leyenda, notar que el borde impide que la leyenda se muestre completamente; E. Reducción del tamaño proporcional del gráfico de pastel para dar chance a que alcance el cuadro de la leyenda.

Gráficos de líneas y puntos

Con la función “plot()” podemos crear diferentes tipos de gráficos únicamente cambiando la letra para cada tipo con el argumento “type=”. Para ejemplificar, utilizaremos los 30 datos de humedad del suelo en tres sitios arreglados por filas, por el momento utilizaremos solamente los valores de la columna HS (los datos completos se muestran en anexo 1):

```
> HS <-read.csv(file.choose())
> head(HS)
  Sitios  HS
1 Sitio1 85.4
2 Sitio1 91.2
3 Sitio1 93.4
4 Sitio1 84.3
5 Sitio1 86.5
6 Sitio1 98.2
> tail(HS)
  Sitios  HS
25 Sitio3 85.8
26 Sitio3 97.5
27 Sitio3 93.6
28 Sitio3 76.5
29 Sitio3 95.4
30 Sitio3 82.5
```

La función “plot()” genera un gráfico de punto por defecto, correspondiente a “type=p” (Figura 105 A):

Aplicaciones de Estadística Básica

```
> plot(HS$HS)
```

Para hacer el gráfico de ejemplo un poco más estético, añadiremos los títulos de los ejes X y Y (Figura 105 B):

```
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)")
```

Entonces utilizamos la función “type=” para cambiar el tipo de gráfico, según el cuadro 19.

Cuadro 19. Opciones para elaborar diferentes tipos de gráficos utilizando la función “plot()” y el argumento “type=”.

TIPO	FUNCIÓN	EJEMPLO
p	Gráfico de puntos (por defecto si no se establece el tipo).	Figura 105 B
l	Gráfico de líneas.	Figura 105 C
b	Gráfico de líneas y puntos con relleno blanco.	Figura 105 D
c	Igual a “b” pero dejando espacio en blanco donde estaban los círculos.	Figura 105 E
o	Gráfico de líneas y puntos con relleno transparente.	Figura 105 F
h	Parecido a histograma, pero con líneas verticales en lugar de barras.	Figura 105 G
s	Parecido a histograma, pero solamente la silueta.	Figura 105 H

A continuación presentamos los comandos para cambiar los tipos de gráficos, utilizando como base el gráfico B de la figura 105, los resultados se muestran en las figuras 105 C - H:

```
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="l")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="b")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="c")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="o")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="h")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="s")
```

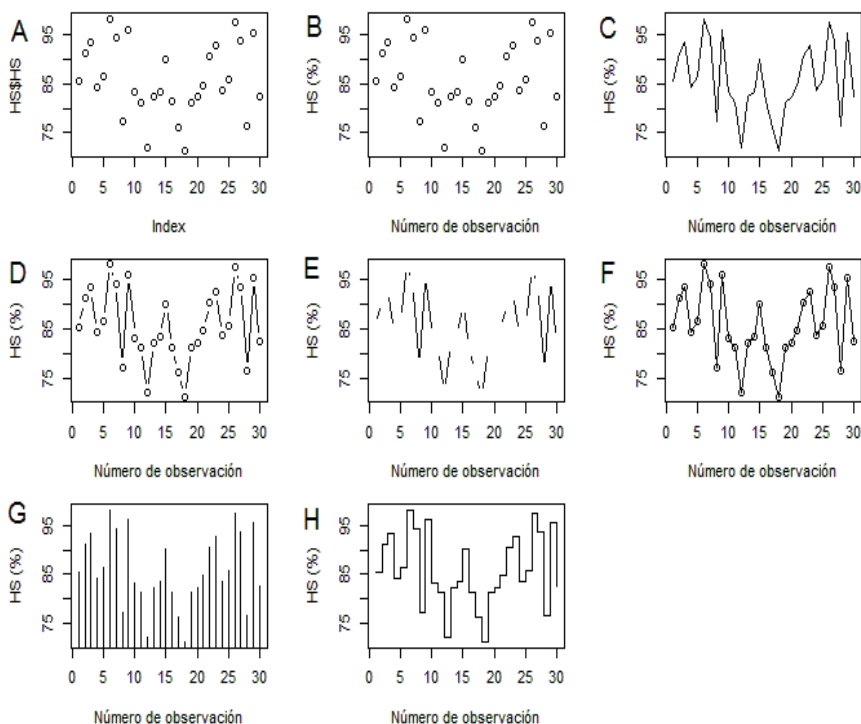



Figura 105. Diferentes tipos de gráficos creados con la función “plot()” y el argumento “type=”, ver cuadro 19.

El tipo de borde de estos gráficos los podemos controlar mediante la función “par()” y el argumento “bty=”. Los valores que se le asignan al argumento “bty=” son “o”, “l”, “7”, “c” o “j”, el significado de cada uno se expresa en el cuadro 20.

Cuadro 20. Opciones para controlar los bordes de los gráficos utilizando la función “par()” y el argumento “bty=”.

TIPO	FUNCIÓN	EJEMPLO
o	Aparecen todos los lados de la caja.	Figura 106 A
l	Aparece solo el lado izquierdo y el de abajo.	Figura 106 B
7	Muestra el lado superior y derecho.	Figura 106 C
c	Muestra el lado superior, izquierdo y abajo.	Figura 106 D
j	Aparece el lado superior, derecho y abajo.	Figura 106 E

Aplicaciones de Estadística Básica

A continuación presentamos los comandos para cambiar los tipos de bordes, utilizando como base el gráfico H de la figura 105, los resultados se muestran en la figura 106:

```
> par(bty="o")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="s")
> par(bty="l")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="s")
> par(bty="7")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="s")
> par(bty="c")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="s")
> par(bty="j")
> plot(HS$HS, xlab="Número de observación", ylab="HS (%)", type="s")
```

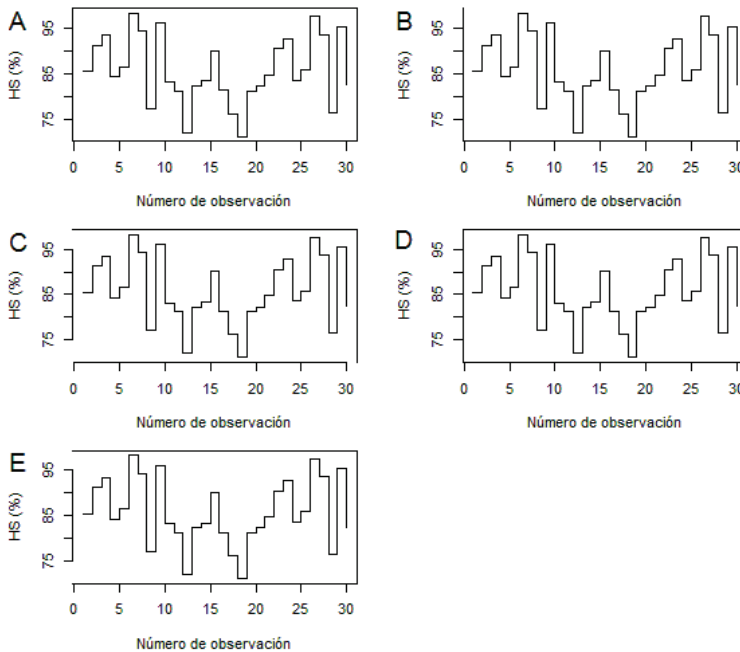


Figura 106. Personalización de los bordes de los gráficos. A. Bordes completos; B. Bordes inferior e izquierdo; C. Bordes superior y derecho; D. Bordes inferior, superior e izquierdo; E. Bordes inferior, superior y derecho.

Con el uso de las funciones “plot()” y “lines()” combinadas, se pueden elaborar gráficos múltiples líneas. Para ejemplificar esta idea, se hará uso de los datos de Oxígeno Disuelto, tomado en diferentes Sitios (fuentes de agua) y en tres meses, los cuales se almacenarán en la variable “OD”:

```
> OD <-read.csv(file.choose())  
> OD
```

	Sitios	Abr	May	Jun
1	1	5.6	4.5	4.7
2	2	4.2	5.5	6.1
3	3	3.1	4.6	4.6
4	4	3.3	4.6	4.8
5	5	6.3	5.4	5.6
6	6	3.5	5.2	6.0
7	7	2.3	3.6	3.8
8	8	3.2	3.2	4.8
9	9	4.1	5.4	5.6
10	10	8.2	2.3	3.9

El gráfico lo construiremos en capas, primero elaboraremos un gráfico con una línea, correspondiente a los datos del mes de abril (Abr) en la tabla de datos (OD). En el argumento “type=” le indicaremos al programa que el gráfico es de línea (l), adicionalmente añadimos las leyendas de los ejes X y Y (Figura 107 A):

```
> plot(OD$Abr ~ OD$Sitios, type="l", xlab="Sitios", ylab="OD (ppm)")
```

Seguidamente agregamos la segunda línea utilizando la función “lines()”, correspondiente a la línea de los datos colectados en el mes de mayo (May) en función de los diez sitios, además del argumento “type=”, le adicionamos el argumento “lty=” que le indica al programa el tipo de línea (sólida, con guiones, punteada, etc.), como ejemplo vamos a usar diferentes tipos de líneas para diferenciarlas (Figura 107 B):

```
> plot(OD$Abr ~ OD$Sitios, type="l", xlab="Sitios", ylab="OD (ppm)")  
> lines(OD$May ~ OD$Sitios, type="l", lty=2)
```

A continuación, añadiremos la tercera línea con la misma función “lines()”, correspondiente a los datos colectados en el mes de junio (Jun) en función de los diez sitios (Figura 107 C):

```
> plot(OD$Abr ~ OD$Sitios, type="l", xlab="Sitios", ylab="OD (ppm)")  
> lines(OD$May ~ OD$Sitios, type="l", lty=2)  
> lines(OD$Jun ~ OD$Sitios, type="l", lty=3)
```

Para completar el gráfico, añadimos la leyenda donde se incluyen los tres tipos de gráficos de líneas (“lty=c(1:3)”) y se especifica el tamaño del cuadro de la leyenda con el argumento “cex=” (Figura 107 D):

Aplicaciones de Estadística Básica

```
> plot(OD$Abr ~ OD$Sitios, type="l", xlab="Sitios", ylab="OD (ppm)")
> lines(OD$May ~ OD$Sitios, type="l", lty=2)
> lines(OD$Jun ~ OD$Sitios, type="l", lty=3)
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3), cex=0.6)
```

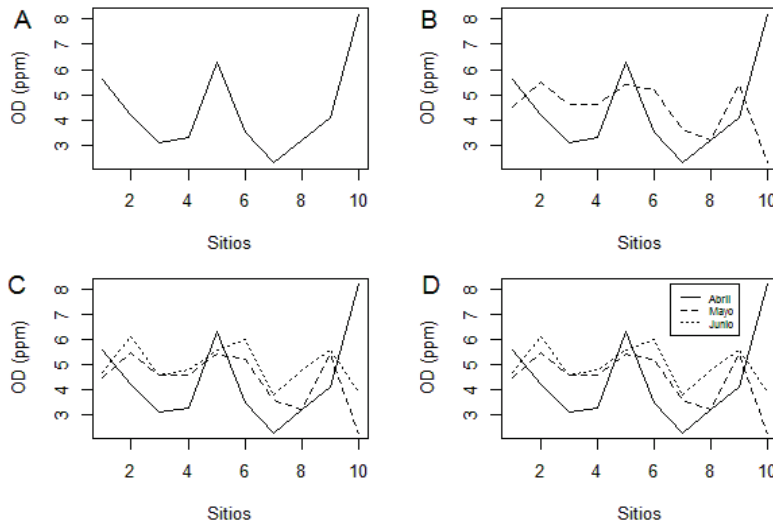


Figura 107. Demostración de elaboración del gráfico de múltiples líneas. A – D explicación del proceso paso a paso.

Podemos asignar colores a las líneas del gráfico de múltiples líneas con el argumento “col=”. Para ejemplificar, retomaremos las líneas de comando utilizadas para crear el gráfico de múltiples líneas anterior y asignaremos los colores azul (blue) para abril, rojo (red) para mayo y verde oscuro (darkgreen) para junio (Figura 108 A):

```
> plot(OD$Abr ~ OD$Sitios, type="l", xlab="Sitios", ylab="OD (ppm)", col="blue")
> lines(OD$May ~ OD$Sitios, type="l", lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="l", lty=3, col="darkgreen")
```

Adicionamos la leyenda, especificando los colores para cada una de las líneas, para ello agregamos de nuevo el argumento “col=” dentro de la función “legend()” y especificamos los colores en un formato vectorial (Figura 108 B):

```
> plot(OD$Abr ~ OD$Sitios, type="l", xlab="Sitios", ylab="OD (ppm)", col="blue")
> lines(OD$May ~ OD$Sitios, type="l", lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="l", lty=3, col="darkgreen")
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3), cex=0.6, col=c("blue", "red", "darkgreen"))
```

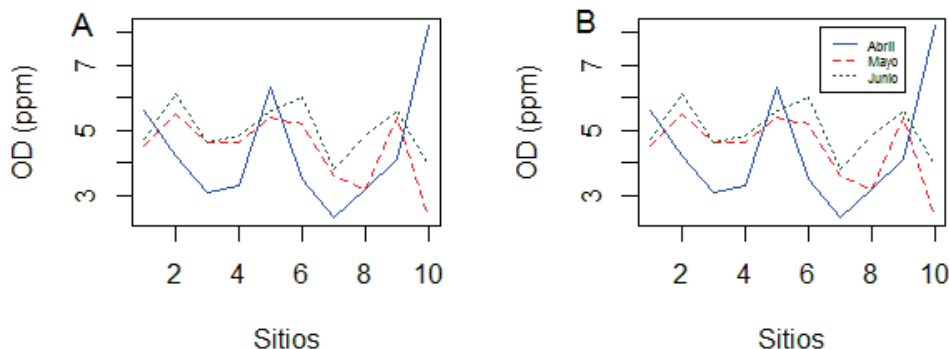


Figura 108. Asignación de colores a cada línea en un gráfico de múltiples líneas. A. Asignación de los colores a cada línea; B. Inserción de la leyenda con los colores y tipos de líneas correspondientes con cada mes.

Aunque no es recomendable, el usuario a su criterio y gusto, puede hacer más complejos los gráficos añadiendo más elementos. Uno de los elementos comúnmente usado en las gráficas de líneas son figuras en sus vértices. Para ellos cambiamos el argumento “type=” de la opción “l” a la opción “o” y nos genera una línea con puntos, de esta manera la forma de los puntos se puede cambiar con el argumento adicional “pch=”. Para ejemplificar, añadiremos un punto cuadrado a mayo (pch=0) y un triángulo invertido a junio (pch=6), abril quedaría con la figura por defecto (pch=1) (Figura 109 A):

```
> plot(OD$Abr ~ OD$Sitios, type="o", xlab="Sitios", ylab="OD
(ppm)", col="blue")
> lines(OD$May ~ OD$Sitios, type="o", pch=0, lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="o", pch=6, lty=3, col=
"darkgreen")
```

Si realizamos estos cambios en el gráfico, estos también se tienen que reflejar en la leyenda; sin embargo, en R, la leyenda no cambia automáticamente con los cambios hechos en el gráfico, de tal forma hay que hacerlo manualmente. Para esto, se añadirá en la función “legend()” el argumento “pch=c(1,0,6)” que le indica al programa las formas que tendrán los puntos en los vértices (Figura 109 B):

```
> plot(OD$Abr ~ OD$Sitios, type="o", xlab="Sitios", ylab="OD
(ppm)", col="blue")
> lines(OD$May ~ OD$Sitios, type="o", pch=0, lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="o", pch=6, lty=3, col="darkgreen")
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3),
pch=c(1,0,6), cex=0.6, col=c("blue", "red", "darkgreen"))
```

Aplicaciones de Estadística Básica

En los gráficos generados con la función “plot()” se hace necesario cambiar los límites de los ejes X y Y, para ajustarlo a la figura del gráfico, esto se logra con los argumentos “xlim=” y “ylim=” y los límites se indican mediante un vector de dos números. Por ejemplo, el gráfico creado anteriormente tiene límites X = 1 a 10 y Y = 2.3 a 8.2; sin embargo, asumamos que queremos modificar esos límites a X = 0 a 11 y Y = 0 a 9, los nuevos argumentos de límites quedarían como “xlim=c(0,11)” y “ylim=c(0,9)” (Figura 109 C):

```
> plot(OD$Abr ~ OD$Sitios, type="o", xlab="Sitios", ylab="OD (ppm)", col="blue", xlim=c(0,11), ylim=c(0,9))
> lines(OD$May ~ OD$Sitios, type="o", pch=0, lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="o", pch=6, lty=3, col="darkgreen")
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3), pch=c(1,0,6), cex=0.6, col=c("blue", "red", "darkgreen"))
```

A gusto del usuario, la leyenda también se puede insertar de forma horizontal, al incluir el argumento “horiz=TRUE” en la función “legend()” (Figura 109 D):

```
> plot(OD$Abr ~ OD$Sitios, type="o", xlab="Sitios", ylab="OD (ppm)", col="blue")
> lines(OD$May ~ OD$Sitios, type="o", pch=0, lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="o", pch=6, lty=3, col="darkgreen")
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3), pch=c(1,0,6), cex=0.6, col=c("blue", "red", "darkgreen"), horiz=TRUE)
```

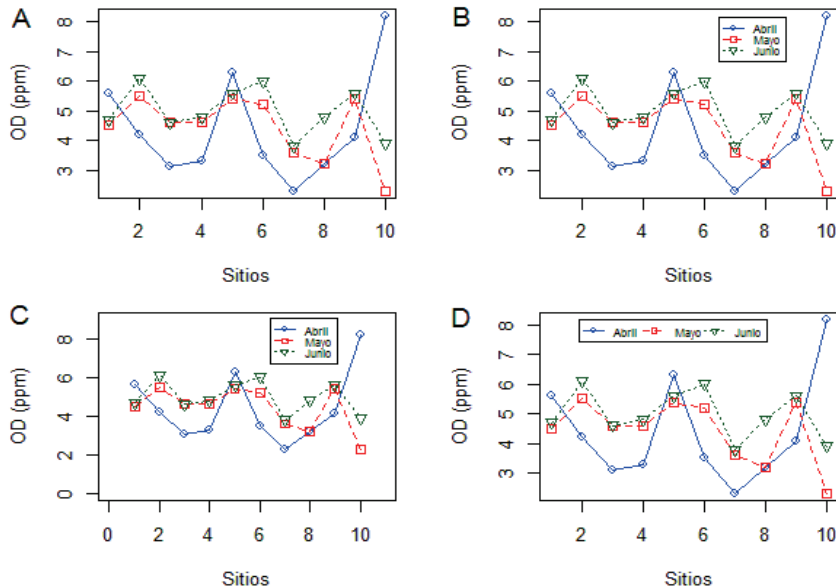


Figura 109. Gráfico de múltiples líneas con puntos de diferentes formas en los vértices. A. Líneas de puntos con diferentes formas (círculo, cuadrado y triángulo invertido, todos huecos); B. Asignación de leyenda; C. Personalización de los límites de los ejes X y Y; D. Personalización de la leyenda a una forma horizontal. El código de los símbolos se visualiza escribiendo “pch” (sin las comillas) en la consola de R.

El recuadro de la leyenda lo podemos personalizar mediante algunos argumentos adicionales incluidos en la función “legend()”, cambiando el tipo, color y grosor del borde. El tipo de borde se controla con el argumento “box.lty=”, cuyas opciones son iguales a las del argumento “lty=”. Si no es de interés mostrar el borde del cuadro de la leyenda se utiliza la opción 0 en el argumento “box.lty=” (Figura 110 A):

```
> plot(OD$Abr ~ OD$Sitios, type="o", xlab="Sitios",
ylab="OD(ppm)", col="blue")
> lines(OD$May ~ OD$Sitios, type="o", pch=0, lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="o", pch=6, lty=3, col=
"darkgreen")
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3),
pch=c(1,0,6), cex=0.6, col=c("blue", "red", "darkgreen"),
horiz=TRUE, box.lty=0)
```

También podemos controlar el color del borde por medio del argumento “box.col=” dentro de la función “legend()” y se seleccionan los colores utilizando los códigos, según el cuadro 18 (Figura 110 B):

```
> plot(OD$Abr ~ OD$Sitios, type="o", xlab="Sitios",
ylab="OD(ppm)", col="blue")
> lines(OD$May ~ OD$Sitios, type="o", pch=0, lty=2, col="red")
> lines(OD$Jun ~ OD$Sitios, type="o", pch=6, lty=3, col="darkgreen")
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3),
pch=c(1,0,6), cex=0.6, col=c("blue", "red", "darkgreen"), horiz=TRUE,
box.col="skyblue")
```

El grosor del borde del cuadro de la leyenda lo personalizamos con el argumento “box.lwd=”, el grosor por defecto es 1, proporcionalmente se reduce o aumenta el grosor, para el ejemplo vamos a establecerlo en un grosor igual a 2 (Figura 110 C):

```
> plot(OD$Abr~OD$Sitios, type="o", xlab="Sitios",
ylab="OD(ppm)", col="blue")
> lines(OD$May~OD$Sitios, type="o", pch=0, lty=2, col="red")
```

Aplicaciones de Estadística Básica

```
> lines(OD$Jun~OD$Sitios, type="o", pch=6, lty=3, col="darkgreen")  
> legend(locator(1), c("Abril", "Mayo", "Junio"), lty=c(1:3),  
pch=c(1,0,6), cex=0.6, col=c("blue","red","darkgreen"), horiz=TRUE,  
box.col="skyblue", box.lwd=2)
```

Solo para ejemplificar, utilizaremos otro tipo de borde, en este caso el borde con guiones, entonces hacemos ese cambio con el uso del argumento "box.lty=2" (Figura 110 D):

```
> plot(OD$Abr ~ OD$Sitios, type="o", xlab="Sitios", ylab="OD (ppm)",  
col="blue")  
> lines(OD$May ~ OD$Sitios, type="o", pch=0, lty=2, col="red")  
> lines(OD$Jun ~ OD$Sitios, type="o", pch=6, lty=3, col="darkgreen")  
> legend(locator(1), c("Abril","Mayo","Junio"), lty=c(1:3),  
pch=c(1,0,6), cex=0.6, col=c("blue","red","darkgreen"), horiz=TRUE,  
box.col="skyblue", box.lwd=2, box.lty=2)
```

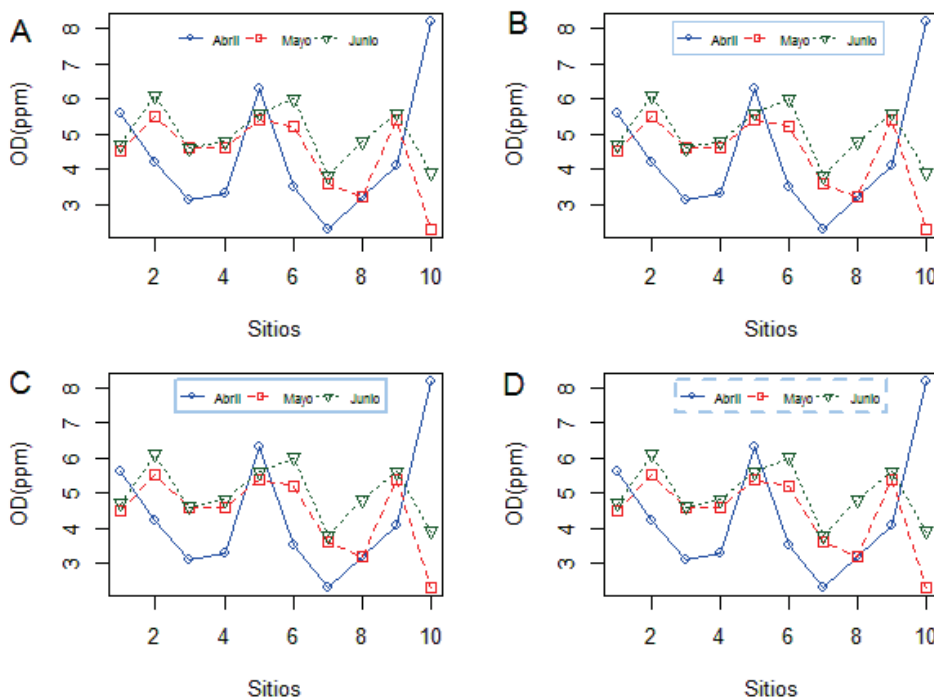


Figura 110. Ilustración de la personalización del borde del cuadro de la leyenda. A. Cuadro de la leyenda sin borde; B. Borde de color celeste; C. Borde más grueso; D. Línea del borde de tipo guiones.

Los gráficos de puntos son muy utilizados para la representación de información. Al igual que los gráficos de líneas, hay muchas formas de personalizarlos. Para ejemplificarlo, utilizaremos los datos de la cobertura de una especie de briófito hepático llamado *Porella platyphylla* y su relación con la temperatura del aire en tres ecosistemas, un ecosistema boscoso (Bosque), un ecosistema agrícola (Agrícola) y un ecosistema urbano (Urbano), los datos los importamos a R y los guardamos en una variable a la que llamaremos "Porella" (los datos se presentan en el anexo 6):

```
> Porella <-read.csv(file.choose())
> head(Porella)
  Arbol      Ecosis Cober Temp
1      1      Bosque  98.1 10.2
2      2      Bosque  86.8 12.4
3      3      Bosque  88.8 10.2
4      4      Bosque  76.9 15.8
5      5      Bosque  83.0 14.5
6      6      Bosque  98.1  9.9
> unique(Porella$Ecosis)
[1] Bosque  Agrícola Urbano
Levels: Agrícola Bosque Urbano
```

Primeramente creamos un gráfico de la relación entre la temperatura y la cobertura del briófito y colocamos las leyendas de los ejes X y Y. En las leyendas se pueden agregar símbolos propios de diferentes medidas. Los superíndices se asignan con el símbolo "[^]", para expresar un valor cualquiera "V" al cuadrado se escribiría "V²" y equivaldría a V²; los subíndices se asignan poniendo entre corchetes el valor del subíndice, por ejemplo para expresar V₂ en el título del eje se representaría como "V[2]".

Las expresiones más complejas se asignan por medio de la función "expression()" dentro del argumento "xlab=" o "ylab=" para representar la leyenda "Temperatura °C", dentro de la expresión se escribiría como "'Temperatura' ~degree~C". Más opciones de expresiones usadas en fórmulas las podemos obtener de <http://vis.supstat.com/2003/04/mathematical-annotation-in-r/>

Con lo anterior en mente, podemos insertar un gráfico de puntos, por defecto los puntos en la función "plot()" son redondos y huecos (Figura 111 A), a partir de ello se pueden personalizar con el argumento correspondiente:

```
> plot(Porella$Temp, Porella$Cober, xlab=expression("Temperatura"
~degree~C), ylab="Cobertura (%)")
```

Aplicaciones de Estadística Básica

La forma del punto la podemos cambiar con el argumento “pch=”, para este ejemplo lo cambiaremos a triángulos invertidos (Figura 111 B):

```
> plot(Porella$Temp, Porella$Cober, xlab=expression("Temperatura"  
~degree~C), ylab="Cobertura (%)", pch=6)
```

El color lo asignamos con el argumento “col=” y las opciones de colores están descritas en el cuadro 18. Para este ejemplo, cambiaremos el color a azul (blue) (Figura 111 C):

```
> plot(Porella$Temp, Porella$Cober, xlab=expression("Temperatura"  
~degree~C), ylab="Cobertura (%)", pch=6, col="blue")
```

El tamaño de cada punto lo controlamos con el argumento “cex=". El valor por defecto es 1, a partir de este valor se asignan valores menores o mayores que reducirán o incrementarán el tamaño de los puntos respectivamente. Para este ejemplo, aumentaremos el tamaño 50% del valor por defecto mediante el argumento “cex=1.5” (Figura 111 D):

```
> plot(Porella$Temp, Porella$Cober, xlab=expression("Temperatura"  
~degree~C), ylab="Cobertura (%)", pch=6, col="blue", cex=1.5)
```

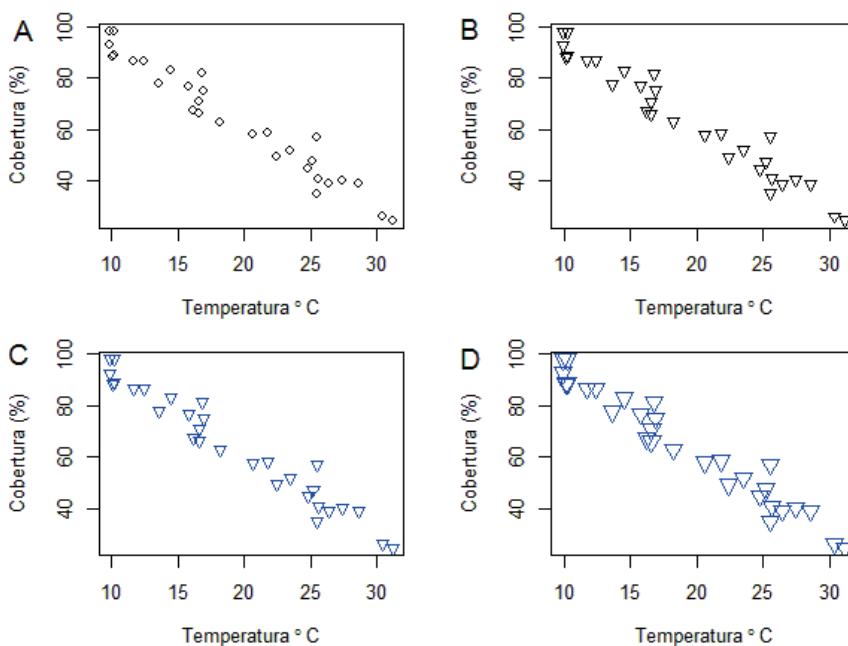


Figura 111. Personalización de los puntos en un gráfico de puntos. A. Tipo, forma, tamaño y color por defecto; B. Cambio a puntos con forma de triángulo invertido; C. Ilustración del cambio en el color de los puntos; D. Ilustración del cambio en el tamaño.

En ciertos casos el usuario está interesado en determinar la relación entre dos variables en función de otra variable categórica que tiene varios niveles. Por ejemplo, es de interés visualizar en un mismo gráfico la relación entre las variables cobertura del briófito (Cober), con la temperatura (Temp), para cada uno de los ecosistemas (Bosque, Agrícola y Urbano).

En este escrito utilizaremos dos formas de alcanzar este objetivo, la primera utilizando la función “with()” y otra utilizando las funciones plot() y lines(). La diferencia entre ambas formas es que con “with()” se utiliza menos codificación, pero los gráficos resultantes no son personalizables; lo contrario caracteriza a la opción utilizando plot() y lines().

En cuanto a la primera opción, dentro de la función “with()” utilizaremos dos cosas, la variable “Porella” donde se encuentran almacenados los datos y la función “plot()”. Dentro de la función “plot()” definiremos las dos variables con las que se construirá el gráfico (Porella\$Temp y Porella\$Cober), y adicionalmente utilizaremos el argumento “pch=” para definir cómo se realizará la separación de los dato, dado a que dicha separación la realizaremos basados en el tipo de ecosistema (Porella\$Ecosis), vamos a transformar los nombres de cada ecosistema en números enteros con el argumento “as.integer()”, de tal forma que el argumento completo quedará descrito como “pch=as.integer(Porella\$Ecosis)”, con el que básicamente le indicamos al programa que ponga símbolos diferentes en dependencia de las categorías de la columna “Ecosis”, dentro de la variable “Porella”.

Los restantes argumentos dentro de la función “plot()” únicamente establecen las leyendas de los ejes X y Y, de tal forma que el comando completo quedaría estructurado de la siguiente manera (Figura 112 A):

```
> with(Porella, plot(Porella$Temp, Porella$Cober, pch=as.integer(
(Porella$Ecosis), xlab=expression("Temperatura" ~degree~C),
ylab="Cobertura (%)"))
```

Notemos que los puntos tienen diferentes formas (símbolos) sobre la base de la variable categórica que indica los tres tipos de ecosistema, triángulo para bosque, círculo para área agrícola y signo de más para urbano. De la misma forma podemos utilizar diferentes colores, en lugar de diferentes símbolos, en cuyo caso solamente se sustituirá el argumento “pch=” por el argumento “col=” en el comando (Figura 112 B):

```
> with(Porella, plot(Porella$Temp, Porella$Cober, col=as.integer(
Porella$Ecosis), xlab=expression("Temperatura" ~degree~C),
ylab="Cobertura (%)"))
```

Aplicaciones de Estadística Básica

Haciéndolo un poco más complejo y vistoso, utilizaremos diferentes tipos de símbolos y colores a la vez, al añadir los dos argumentos “pch=” y “col=” (Figura 112 C):

```
> with(Porella, plot(Porella$Temp, Porella$Cober, pch=as.integer  
(Porella$Ecosis), col=as.integer(Porella$Ecosis), xlab=expression  
("Temperatura" ~degree~C), ylab="Cobertura (%)"))
```

Para agregar la leyenda, añadimos una nueva línea de comando utilizando la función “legend()” y cuatro argumentos principales, el argumento “locator(1)” para tener la opción de colocar la leyenda en cualquier parte del área de gráfico solamente haciendo clic; los nombres de cada categoría dentro de la leyenda mediante un vector; los tipos de símbolos mediante el argumento “pch=”; el color y el tamaño del cuadro de la leyenda (Figura 112 D):

```
> with(Porella, plot(Porella$Temp, Porella$Cober, pch=as.integer  
(Porella$Ecosis), col=as.integer(Porella$Ecosis), xlab=expression  
("Temperatura" ~degree~C), ylab="Cobertura (%)"))  
> legend(locator(1), c("Bosque", "Agricola", "Urbano"), pch=c(2, 1, 3),  
col=c("red", "black", "green"), cex=0.6)
```

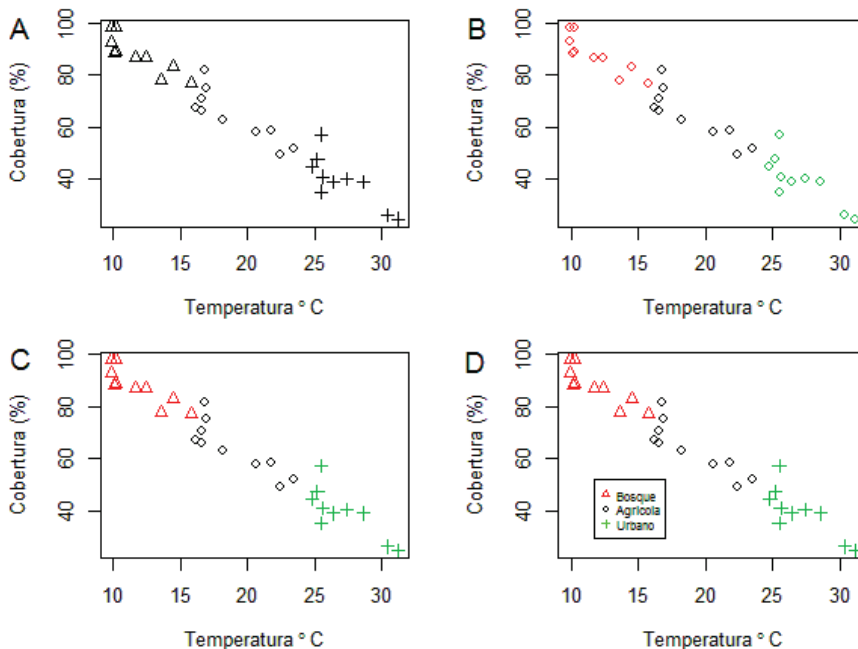


Figura 112. Gráfico de puntos en los cuales se utilizan diferentes formas (símbolos) y colores basados en la variable categórica “Ecosis” (ecosistema) con sus tres categorías bosque, área agrícola y área urbana, utilizando la función “with()”. A – C. Demostración de la personalización del tipo de punto y el color; D. Asignación de la leyenda.

La opción anterior no nos permite controlar las formas y colores de dichos puntos, pues el programa los asigna por defecto, de tal forma que para poder controlar y establecer las formas y colores de manera más personalizada. Tendríamos que hacer uso de la segunda opción, la cual es un poco más tediosa en términos de pasos y codificaciones.

Elaboraremos el gráfico por capas, para ello es necesario desde la primera capa establecer los límites de los ejes X y Y de toda el área de gráfico, esto lo logramos conociendo los valores mínimos y máximos para cada variable a graficar. Utilizaremos la función “sapply()” con el argumento “FUN=min” y “FUN=max” para determinar los valores mínimos y máximos de las dos variables a usar en el gráfico (Cober y Temp):

```
> sapply(Porella[,3:4], FUN=min)
Cober  Temp
 24.6   9.9
> sapply(Porella[,3:4], FUN=max)
Cober  Temp
 98.1  31.2
```

Con los valores mínimos y máximos podemos establecer los límites, en este caso estableceremos los límites entre 9 y 32 en el eje X, donde se desplegará la variable temperatura (Temp) y a 23 y 100 el eje Y donde se mostrará la variable cobertura (Cober). Luego incertamos el primer gráfico con la función “plot()” y los dos argumentos iniciales le indican al programa que deseamos elaborar un gráfico de punto de las variables temperatura (Temp) y cobertura (Cober) solamente del set de datos correspondiente a la categoría de bosque, dentro de la columna “Ecosis”. En la figura 113 se representa el trabajo que ejecuta este primer argumento.

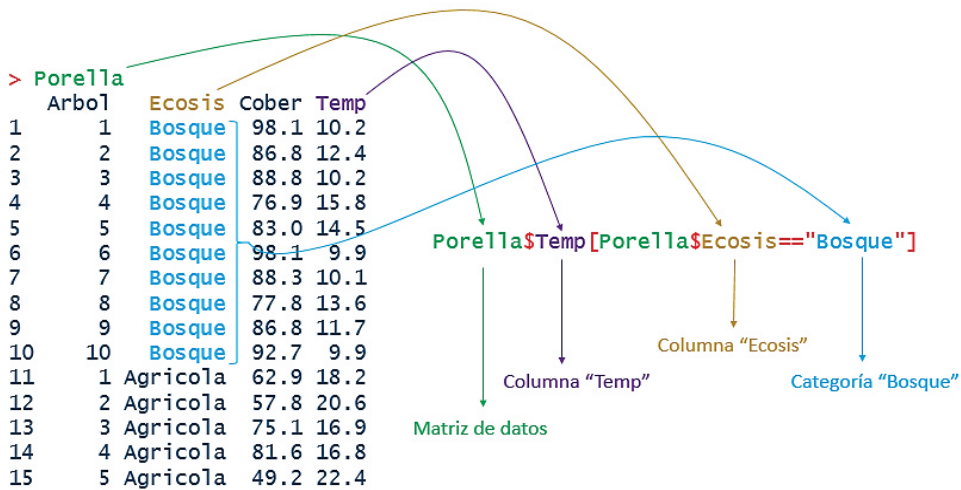


Figura 113. Ilustración de argument “Porella\$Temp[Porella\$Ecosis=="Bosque"]” con el cual básicamente se le indica al programa que utilice los datos de la variable temperatura, que corresponda solamente a la categoría de bosque (subconjunto) dentro de la variable en la columna “Ecosis”.

El primer gráfico elaborado con el subconjunto de datos según la categoría “Bosque” sería construido con el comando (Figura 114 A):

```
> plot(Porella$Temp[Porella$Ecosis=="Bosque"], Porella$Cober
[Porella$Ecosis=="Bosque"], xlab=expression("Temperatura"
~degree~C), ylab="Cobertura (%)", pch=16, col="blue", xlim=c(9,32),
ylim=c(23,100))
```

Notemos que los argumentos “xlim=” y “ylim=” establecen los límites de los ejes X y Y respectivamente. Adicionalmente se asignan puntos sólidos (pch=16) de color azul (blue).

Seguidamente se agrega la segunda capa de puntos utilizando la función “points()” y como primeros dos argumentos se definen los datos a utilizar mediante el subconjunto de datos según la categoría “Agrícola”. Además se asignan puntos sólidos (pch=16) de color rojo (red) (Figura 114 B):

```
> plot(Porella$Temp[Porella$Ecosis=="Bosque"],
Porella$Cober[Porella$Ecosis=="Bosque"], xlab=expression
("Temperatura"~degree~C), ylab="Cobertura (%)", pch=16,
col="blue", xlim=c(9,32), ylim=c(23,100))
> points(Porella$Temp[Porella$Ecosis=="Agrícola"], Porella$
```

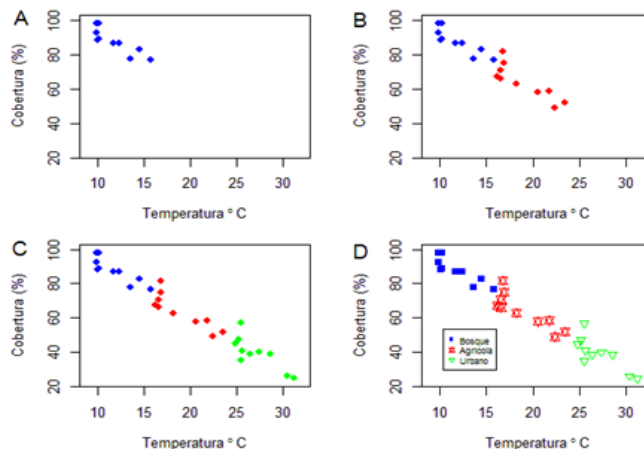
```
Cober[Porella$Ecosis=="Agricola"], pch=16, col="red")
```

Finalmente, así como añadimos los puntos para la categoría “Agricola” de la variable “Ecosis”, se añaden los puntos de la categoría “Urbano”, cambiando el color a verde (green) (Figura 114 C):

```
> plot(Porella$Temp[Porella$Ecosis=="Bosque"],  
Porella$Cober[Porella$Ecosis=="Bosque"], xlab=expression(  
("Temperatura"~degree~C), ylab="Cobertura (%)", pch=16,  
col="blue", xlim=c(9,32), ylim=c(23,100))  
> points(Porella$Temp[Porella$Ecosis=="Agricola"],  
Porella$Cober[Porella$Ecosis=="Agricola"], pch=16, col="red")  
> points(Porella$Temp[Porella$Ecosis=="Urbano"],  
Porella$Cober[Porella$Ecosis=="Urbano"], pch=16, col="green")
```

Además de controlar los colores de los puntos, también podemos controlar el tipo de punto añadiendo el argumento “pch=”. Adicionamos la leyenda al final de estas líneas de comando (Figura 114 D):

```
> plot(Porella$Temp[Porella$Ecosis=="Bosque"], Porella$Cober  
[Porella$Ecosis=="Bosque"], xlab=expression("Temperatura"  
~degree~C), ylab="Cobertura (%)", pch=15, col="blue", xlim=c(9,32),  
ylim=c(23,100))  
> points(Porella$Temp[Porella$Ecosis=="Agricola"],  
Porella$Cober[Porella$Ecosis=="Agricola"], pch=11, col="red")  
> points(Porella$Temp[Porella$Ecosis=="Urbano"],  
Porella$Cober[Porella$Ecosis=="Urbano"], pch=6, col="green")  
> legend(locator(1), c("Bosque", "Agricola", "Urbano"), pch=c(15,11,6),  
col=c("blue", "red", "green"), cex=0.6)
```



Aplicaciones de Estadística Básica

Figura 114. Gráfico de puntos en los cuales se utilizan diferentes formas (símbolos) y colores basados en la variable categórica “Ecosis” (ecosistema) con sus tres categorías bosque, área agrícola y área urbana, utilizando las funciones “plot()” y “points()”. A – C. Demostración de la personalización del tipo de punto y el color; D. Asignación de la leyenda.

Matriz de gráficos de punto

Este es un gráfico muy popular para determinar relaciones entre más de dos variables, tiene un formato matricial y el gráfico correspondiente a las variables se determina al hacer coincidir cada par de variables. La diagonal del cuadrado está ocupada con el nombre de las variables y la información que se despliega en la parte de arriba de dicha diagonal, es igual a la de la parte de abajo.

Para ilustrar la elaboración de este gráfico utilizaremos una base de datos del briófito hepático *Porella platyphylla*, pero incluyendo dos variables más: humedad relativa del aire (HR) y concentración (ppm) de CO₂ (CO2). A continuación se observa la estructura de la tabla de datos guardados en la variable “Porella2” (base completa en anexo 7):

```
> Porella2 <-read.csv(file.choose())
> head(Porella2)
  Arbol  Sitio Cober Temp   HR   CO2
1     1   Bosque  98.1 10.2  92.5 316.4
2     2   Bosque  86.8 12.4  92.4 313.0
3     3   Bosque  88.8 10.2 104.0 304.9
4     4   Bosque  76.9 15.8  89.0 342.6
5     5   Bosque  83.0 14.5  75.7 322.8
6     6   Bosque  98.1  9.9  91.6 300.2
```

Utilizaremos la función “plot()” o “pairs()” y como argumento especificaremos las columnas que contienen las variables numéricas, que serán usadas en el gráfico, estas serían de la columna 3 a la 6 (3:6) (Figura 115):

```
> plot(Porella2[,3:6])
```

Este tipo de gráfico también permite personalización, para demostrarlos vamos a reescribir los nombres de las variables en la diagonal con el argumento “labels=”, dichos nombres los presentaremos en negritas con el argumento “font.labels=”, reduciremos el tamaño de los mismos con el argumento “cex.labels=”, controlaremos la distancia de separación de cada cuadro dentro del gráfico con “gap=” y quitaremos el panel que está debajo de la diagonal con “lower.panel=” más el argumento “NULL” (la opción contraria es “upper.panel=”) (Figura 115).


```
> plot(Porella2[,3:6], labels=c("Cobertura", "Temperatura",  
"Humedad Relativa", "Dióxido de Carbono"), font.labels=2, cex.  
labels=0.9, gap=0, lower.panel=NULL)
```

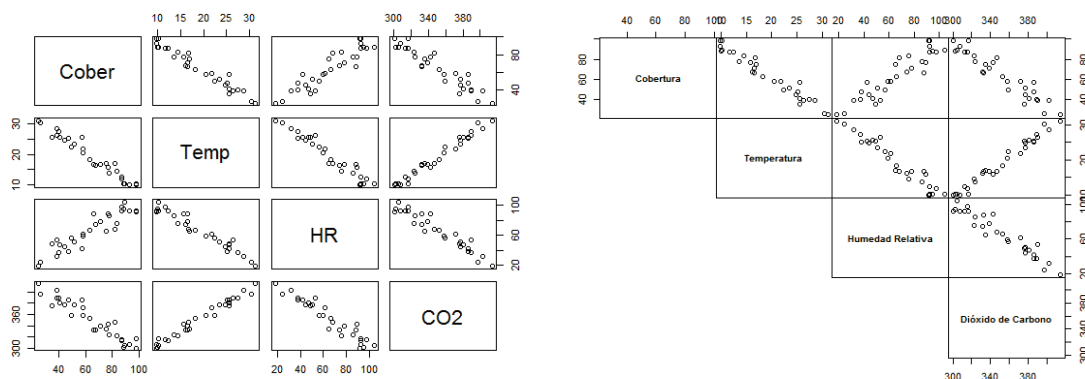


Figura 115. Matriz de gráfico de puntos utilizando cuatro variables: Cober= Cobertura de la especie *P. platyphylla*; Temp= Temperatura del aire; HR = Humedad relativa del aire y CO₂= Concentración (ppm) de CO₂. A la izquierda se presenta la matriz por defecto, a la derecha, se presenta la matriz personalizada.

Las barras de error

Las barras de error en los gráficos son muy comunes y útiles, en especial para obtener una idea general del tamaño del error y su relación con otras barras de error, en el mismo gráfico a fin de examinar (de manera general) diferencias significativas (Figura 68). En las opciones por defecto de R las barras de error se integran en los gráficos después de cumplido una serie de pasos y codificación que a continuación vamos a practicar; sin embargo, es bueno decir hay paquetes gráficos que se pueden instalar en R y generan las barras de error de una forma automática o casi automática.

Para ejemplificar la construcción de un gráfico de barra con barras de error, utilizaremos mediciones de humedad relativa del aire tomadas en cinco puntos, en cada uno de tres sitios. Importamos y guardamos los datos en la variable "HR":

```
> HR <-read.csv(file.choose())  
> HR  
  Sitios  HR  
1  Sitio1 62.3  
2  Sitio1 91.2  
3  Sitio1 56.9
```

Aplicaciones de Estadística Básica

```
4 Sitio1 84.3
5 Sitio1 22.1
6 Sitio2 81.2
7 Sitio2 72.1
8 Sitio2 77.3
9 Sitio2 83.4
10 Sitio2 60.6
11 Sitio3 84.7
12 Sitio3 90.5
13 Sitio3 92.7
14 Sitio3 83.6
15 Sitio3 62.1
```

En primer lugar, calcularemos todos los valores que necesitaremos para elaborar, tanto el gráfico de barras como las barras de error, esto son: media, desviación estándar (DE), número de observaciones (N) y error estándar (EE). Para los primeros tres valores utilizaremos la función “`tapply()`” e iremos cambiando la función estadística a utilizar según corresponda (“`mean`” para la media; “`sd`” para la desviación estándar y “`length`” para el número de observaciones). Los dos argumentos iniciales de la función “`tapply()`” son la columna donde están los valores que se usarán en el cálculo (HR\$HR) y la columna donde está la variable categórica, en función de la cual se harán los cálculos (HR\$Sitios); el tercer argumento es la operación estadística (ver notas precedidas con el símbolo “#”):

```
> Medias <-tapply(HR$HR, HR$Sitios, mean) # Calcula las medias
> Medias
Sitio1 Sitio2 Sitio3
 63.36  74.92  82.72
> DE <-tapply(HR$HR, HR$Sitios, sd) # Calcula las desviaciones
estándares
> DE
      Sitio1      Sitio2      Sitio3
27.197390  9.083336 12.145040
> N <-tapply(HR$HR, HR$Sitios, length) # Calcula el número de
observaciones
> N
Sitio1 Sitio2 Sitio3
     5     5     5
```

El error estándar (EE) lo calcularemos a modo de fórmula, pues no hay ninguna fórmula o argumento para calcularlo, de tal forma que utilizaremos la fórmula “ DE/\sqrt{N} ” o sea la desviación estándar dividida entre la raíz cuadrada del número de observaciones. Los resultados se guardan en la variable “EE”:

```
> EE <-DE/sqrt(N) # Calcula el error estándar
> EE
      Sitio1      Sitio2      Sitio3
12.163042   4.062192   5.431427
```

Luego de los cálculos anteriores procedemos a crear el gráfico de barra y lo guardamos en una variable, en este caso llamada “Grafico” y a la vez asignamos las leyendas de los ejes X y Y (Figura 116 A):

```
> Grafico <-barplot(Medias, xlab="Sitios", ylab="Humedad
Relativa") # Crea el gráfico de barras
```

Seguidamente, agregamos las barras de error con el uso de la función “arrows()” y como argumentos dentro de esta variable le indicamos al programa la variable donde está guardado el gráfico (“Grafico”) y las medias; de nuevo escribimos el nombre de la variable donde está guardado el gráfico y las medias más el error; finalmente le indicamos que las barras estarán a 90 grados con respecto a las medias (Figura 116 B):

```
> arrows(Grafico, Medias, Grafico, Medias+EE, angle=90) # Agrega
barra de error positiva
```

Anteriormente, agregamos la barra de error positiva (hacia arriba), pero también podemos agregar la barra de error negativa (hacia abajo). Esto es sencillo, pues solo se agrega otra capa en el gráfico con el uso de una tercera línea de comandos que contenga la función “arrows()”, pero que el argumento “Medias+EE” se cambia a “Medias-EE” (Figura 116 C):

```
> Grafico <-barplot(Medias, xlab="Sitios", ylab="Humedad Relativa")
> arrows(Grafico, Medias, Grafico, Medias+EE, angle=90)
> arrows(Grafico, Medias, Grafico, Medias-EE, angle=90)
```

El grosor de las líneas que conforman las barras de error se puede controlar con el argumento “lwd=”, mientras que la longitud de la barra horizontal de la barra de error se puede controlar con el argumento “length=” (Figura 116 D):

```
> Grafico <- barplot(Medias, xlab="Sitios", ylab="Humedad Relativa")
> arrows(Grafico, Medias, Grafico, Medias+EE, angle=90, lwd=2, length=0.1)
> arrows(Grafico, Medias, Grafico, Medias-EE, angle=90, lwd=2, length=0.1)
```

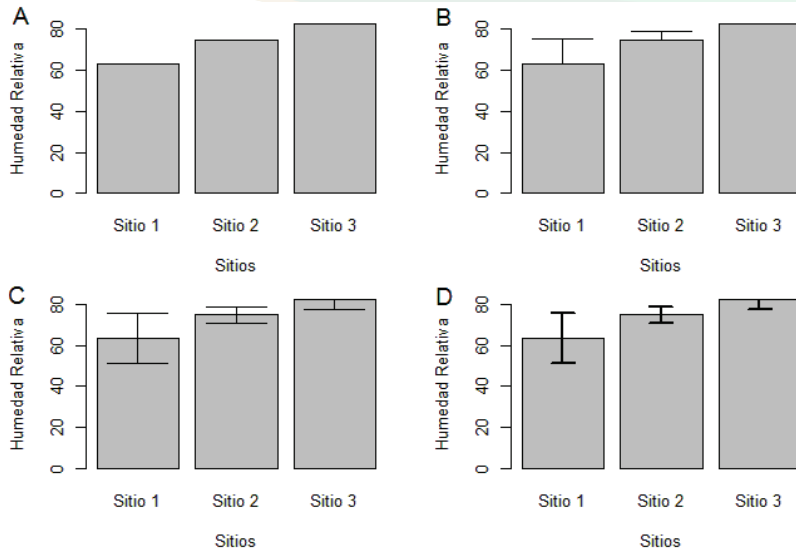


Figura 116. Proceso para agregar barras de error a un gráfico de barras. A. Gráfico de barra sin las barras de error; B. Gráfico de barra con las barras de error positivas (hacia arriba); C. Gráfico de barra con las barras de error positivas y negativas (hacia arriba y hacia abajo); D. Personalización de las barras de error: incremento del grosor y reducción del tamaño de la barra horizontal. Interpretación de los casos C y D en la figura 68.

Las barras de error se suelen agregar a los gráficos de puntos también, para ejemplificarlo utilizaremos los mismos datos anteriores y haremos uso de la función “stripchart()”. Esta función nos permite agregar capas para formar un solo gráfico. Primeramente agregamos los puntos del gráfico con “stripchart()” y dentro de esta función anexamos cinco argumentos, el primero es el que le indica al programa la variable numérica en función de la variable categórica; el segundo le indica que los puntos del gráfico los coloque en posición vertical (“vertical=TRUE”); el tercero le indica el tipo de punto (“pch=1”); el cuarto indica la cantidad de desagregación entre los puntos (“jitter=”); y el quinto argumento designa el título del eje Y (Figura 117 A):

```
> stripchart(HR$HR ~ HR$Sitios, vertical=TRUE, pch=1,
jitter=0.05, xlab="Sitios", ylab="Humedad Relativa")
```

Seguidamente, agregamos los puntos con los valores de las medias, como una capa que aparecerá sobre el stripchart, el argumento “1:3” representa la posición de cada categoría de la variable “Sitios” (Figura 117 B):

```
> stripchart(HR$HR ~ HR$Sitios, vertical=TRUE, pch=1,
jitter=0.05, xlab="Sitios", ylab="Humedad Relativa")
> points(1:3, Medias, pch=16, cex=1.5)
```

Seguidamente se agregan las barras de error positiva y negativa con los argumentos "1:3", "Medias+EE", "angle=" y "length=" (Figura 117 C):

```
> stripchart(HR$HR ~ HR$Sitios, vertical=TRUE, pch=1,  
jitter=0.05, xlab="Sitios", ylab="Humedad Relativa")  
> points(1:3, Medias, pch=16, cex=1.5)  
> arrows(1:3, Medias, 1:3, Medias+EE, angle=90, length=0.08)  
> arrows(1:3, Medias, 1:3, Medias-EE, angle=90, length=0.08)
```

Para personalizar los colores, se asignan los colores con el argumento "col=" dentro de las funciones "points()" y "arrows()" (Figura 117 D):

```
> stripchart(HR$HR ~ HR$Sitios, vertical=TRUE, pch=1, jitter=0.05,  
xlab="Sitios", ylab="Humedad Relativa", col="blue")  
> points(1:3, Medias, pch=16, cex=1.5, col="red")  
> arrows(1:3, Medias, 1:3, Medias-EE, angle=90, length=0.08,  
col="red")  
> arrows(1:3, Medias, 1:3, Medias+EE, angle=90, length=0.08,  
col="red")
```

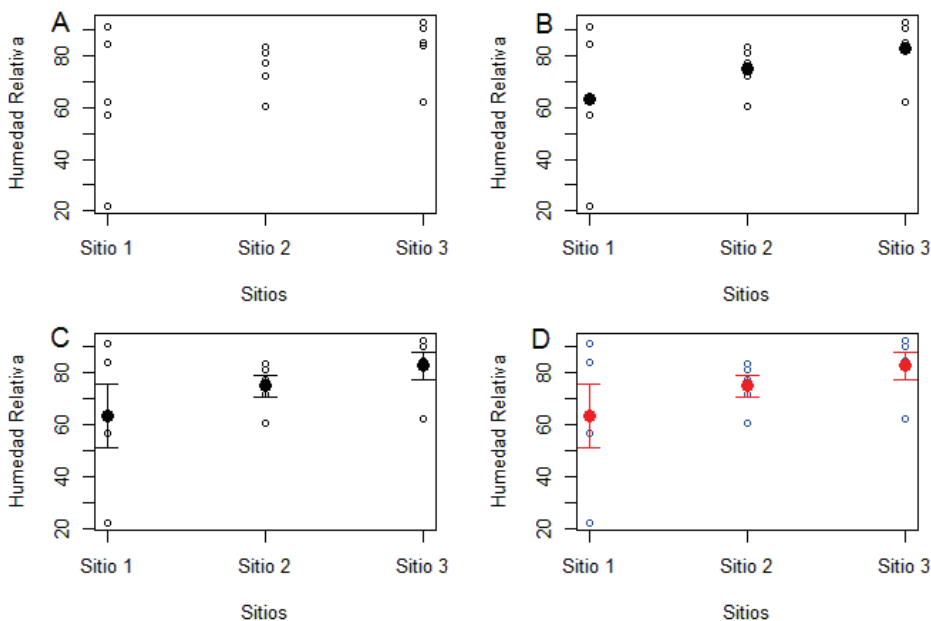


Figura 117. Proceso para agregar barras de error a un gráfico de puntos con la función "stripchart()". A – C. Pasos para agregar las capas que componen el gráfico; D. Personalización de los colores.

Otros gráficos

Gráfico de cajas

Los gráficos de caja son muy populares para visualizar la distribución de los datos y determinar datos atípicos. El valor central que representa el gráfico de caja es la mediana. La parte inferior y superior de la caja está representada por el primero (Q1) y tercer cuartil (Q3), las barras representan la distribución de los datos, excluyendo los valores atípicos y los cuales son definidos por defecto como cualquier valor 1.5 x en rango intercuartil ($Q3 - Q1$) (Teetor, 2011) (Figura 118).

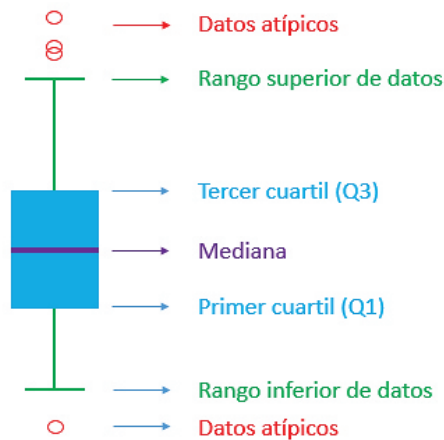


Figura 118. Explicación ilustrativa de las partes de un gráfico de caja.

Utilizaremos los datos de humedad relativa arreglados por filas (utilizado anteriormente) para ejemplificar su uso:

```
> HR <- read.csv(file.choose())
> head(HR)
  Sitios  HR
1 Sitio1 62.3
2 Sitio1 91.2
3 Sitio1 56.9
4 Sitio1 84.3
5 Sitio1 22.1
6 Sitio2 81.2
> unique(HR$Sitios)
[1] Sitio1 Sitio2 Sitio3
Levels: Sitio1 Sitio2 Sitio3
```

Para añadir el gráfico, se utiliza la función “plot()” y un solo argumento que indica los valores numéricos en función de una variable categórica, en este caso los valores de HR en función de los sitios, expresados en el argumento “HR\$HR ~ HR\$Sitios”, adicionalmente se le añaden las leyendas de los ejes X y Y (Figura 119 A):

```
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad  
Relativa (%)")
```

El ancho de la caja la podemos controlar con el argumento “boxwex=". El valor del ancho por defecto es 0.8, el ancho adecuado depende del gusto del usuario, quien puede probar varias veces hasta determinar el ancho que más le guste, para este ejemplo lo dejaré en un ancho de 0.3 (Figura 119 B):

```
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad  
Relativa (%)", boxwex=0.3)
```

El argumento “whisklty=” toma los mismos valores del argumento “lty=” y es útil para personalizar el tipo de línea, que representa la dispersión de los datos. Para ejemplificar, cambiaremos la línea con guiones que aparece por defecto por una línea sólida, anexando el argumento “whisklty=1” (Figura 119 C):

```
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad  
Relativa (%)", boxwex=0.3, whisklty=1)
```

El color de fondo de la caja y de la orilla son controlados por los argumentos “col=” y “border=”, asignando valores codificados, según el cuadro 18 (Figura 119 D):

```
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad  
Relativa (%)", boxwex=0.3, whisklty=1, col="green",  
border="blue")
```

Los gráficos de caja también los podemos disponer de forma horizontal, para ellos se añade a la función “plot()” el argumento “horizontal=TRUE” (Figura 119 E):

```
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad  
Relativa (%)", boxwex=0.3, whisklty=1, col="green",  
border="blue", horizontal=TRUE)
```

El color del cuadro que representa el borde del gráfico, puede ser personalizado con un comando previo utilizando la función “par()” y el argumento “fg=”, en el cual se define el color (Figura 119 F):

Aplicaciones de Estadística Básica

```
> par(fg="red")
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad
Relativa (%)", boxwex=0.3, whisklty=1, col="green",
border="blue", horizontal=TRUE)
```

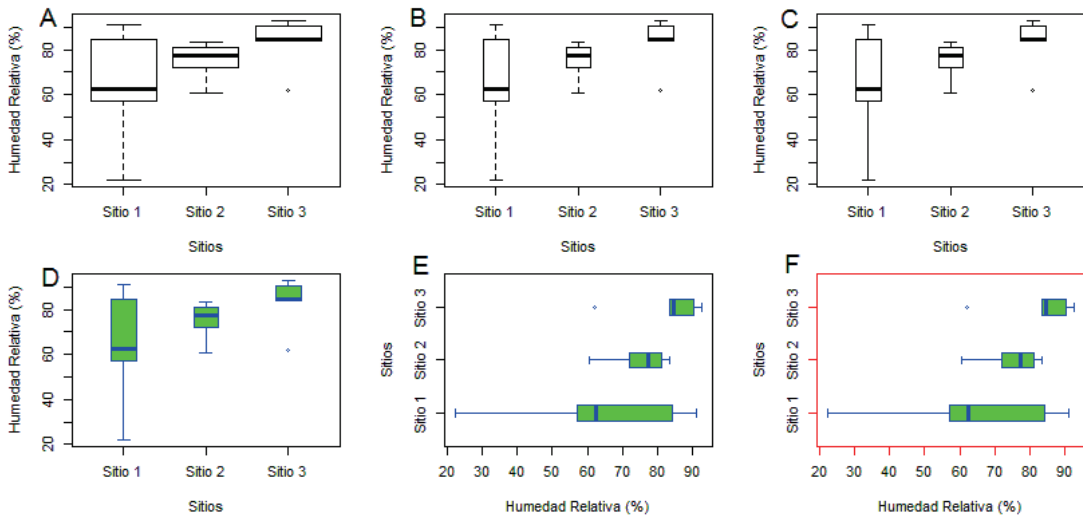


Figura 119. Creación y personalización de gráficos de caja. A. Gráfico por defecto; B. Reducción del ancho de las cajas; C. Cambio de la línea vertical de guiones a sólida; D. Asignación de colores de relleno y de borde de la caja; E. Cambio de las cajas a posición horizontal; F. Personalización del color del borde del gráfico.

Continuando con la personalización de los gráficos de caja, tomaremos el gráfico E de la figura 119 para hacer la demostración de cómo se modifica el eje Y. En principio, vamos a establecer los nombres del eje Y en posición horizontal con el uso del argumento “las=1” (Figura 120 A):

```
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad
Relativa (%)", boxwex=0.3, whisklty=1, col="green",
border="blue", horizontal=TRUE, las=1)
```

Sin embargo, el resultado no es estético, ya que al poner los nombres en posición horizontal, estos se sobrepone con el título del eje Y. Para solucionar el problema se ofrece dos opciones: quitar el nombre de la etiqueta o mover el título del eje Y.

Para quitar el título del eje Y, simplemente establecemos el argumento que denota el eje X como “xlab=NA”, lo hacemos con el eje X, pues el programa sigue considerando al eje Y como eje X tras el cambio de posición de vertical a horizontal (Figura 120 B):


```
> plot(HR$HR ~ HR$Sitios, xlab=NA, ylab="Humedad Relativa (%)",  
boxwex=0.3, whisklty=1, col="green", border="blue",  
horizontal=TRUE, las=1)
```

Si el título del eje Y es absolutamente necesario, entonces utilizaríamos la opción de mover el título del eje Y. Esto lo logramos con el argumento “mgp=” dentro de la función “par()”. El argumento “mgp=” es comandado por un vector de números enteros que denotan el número de líneas donde se encuentran las leyendas de los ejes, los nombres o números de los ejes y la escala del eje respectivamente con respecto al margen del gráfico. Para determinar los valores por defecto del argumento “mgp=” se utiliza el comando “par(“mgp””, este genera tres valores, el primero controla la distancia de las leyendas de los ejes con respecto al borde del gráfico, el segundo controla la distancia de los nombres de los ejes, con respecto al borde del gráfico y el tercero controla la distancia de las líneas de los ejes, con respecto al borde del gráfico:

```
> par("mgp")  
[1] 3 1 0
```

Vamos a aumentar la distancia de las leyendas de los gráficos tornando 4 el número 3 (Figura 120 C):

```
> par(mgp=c(4,1,0))  
> plot(HR$HR ~ HR$Sitios, xlab="Sitios", ylab="Humedad  
Relativa (%)", boxwex=0.3, whisklty=1, col="green",  
border="blue", horizontal=TRUE, las=1)
```

Ahora el título del eje Y (Sitios) se observa separado de los nombres del mismo eje (Sitio1, Sitio2 y Sitio3), sin embargo esto trae consigo otro problema, como hemos aumentado la distancia del eje Y con respecto al borde del gráfico, también ha aumentado de forma simultánea la distancia del eje X con respecto a su borde, por lo tanto tenemos un gráfico poco estético, pues el título del eje X está muy separado del resto del gráfico.

Para resolver esto definitivamente, vamos a quitar el título de eje Y y lo anexaremos con un comando aparte en forma de capa, con la función “title()” y dos argumentos, el primero asigna el nuevo título del eje Y (ylab="Sitios"), el segundo establece el número de la línea donde se colocará la leyenda del eje Y. Después de unas pruebas, he determinado que la mejor posición de la leyenda es en la línea 4 (Figura 120 D):

```
> par(mgp=c(3,1,0))  
> plot(HR$HR ~ HR$Sitios, xlab=NULL, ylab="Humedad Relativa  
(%)", boxwex=0.3, whisklty=1, col="green", border="blue",  
horizontal=TRUE, las=1)  
> title(ylab="Sitios", line=4)
```

Aplicaciones de Estadística Básica

Notemos que también agregamos de nuevo el argumento “mgp=”, pues anteriormente establecimos el primer número en 4, ahora lo tenemos que regresar a 3, su valor por defecto.

Como última característica a personalizar está el color de las cajas. Si se requiere que cada caja tenga diferentes colores en lugar de uno uniforme, entonces hacemos el cambio en el argumento “col=” y establecemos los colores deseados: verde (green), violeta (violet) y celeste profundo (deepskyblue) (Figura 120 E):

```
> plot(HR$HR ~ HR$Sitios, xlab=NULL, ylab="Humedad Relativa (%)", boxwex=0.3, whisklty=1, col=c("green","violet", "deepskyblue"), border="blue", horizontal=TRUE, las=1)
> title(ylab="Sitios", line=4)
```

También se pueden asignar colores con paletas de colores preestablecidas por R, como la paleta llamada arcoíris (rainbow). Esto lo logramos incluyendo el código “rainbow(length(unique(HR\$Sitios)))” en el argumento “col=”. Notemos que “HR\$Sitios” indica las categorías en las que está dividida la variable “Sitios” (Figura 120 F):

```
> plot(HR$HR ~ HR$Sitios, xlab=NULL, ylab="Humedad Relativa (%)", boxwex=0.3, whisklty=1, col=rainbow(length(unique(HR$Sitios))), border="blue", horizontal=TRUE, las=1)
> title(ylab="Sitios", line=4)
```

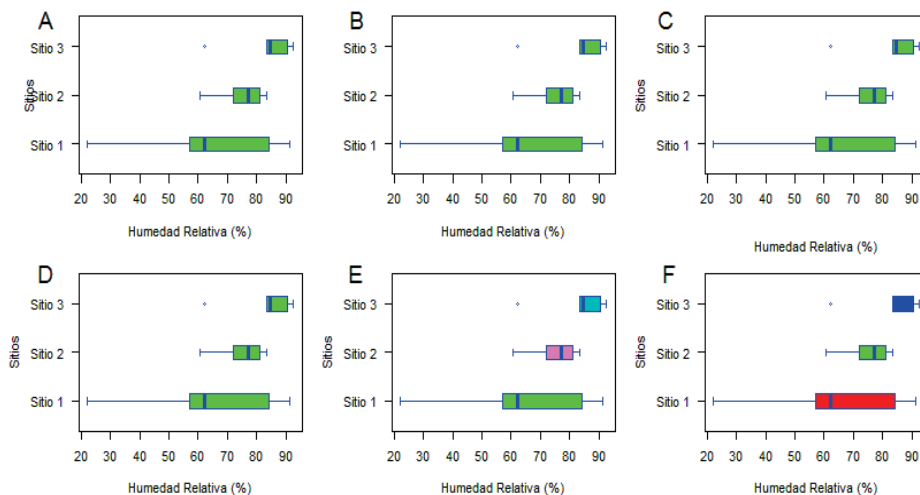


Figura 120. Personalización del eje Y de un gráfico de cajas. A. Título del eje Y (Sitios) se superpone con los nombres del mismo eje (Sitio1, Sitio2, Sitio3). C. Incremento de la distancia entre los títulos de los ejes con respecto al borde del gráfico; D. Asignación del título del eje Y utilizando un comando con la función “title()”; E. Asignación de colores

diferentes a las cajas; F. Asignación de colores a las cajas utilizando una paleta de colores predefinida.

Gráficos de densidad y de violín

Los gráficos de densidad y de violín son ampliamente utilizados para representar información. El gráfico de densidad, específicamente representa datos de una variable numérica de una forma muy parecida a la presentada por un histograma, pero con líneas suaves, en lugar de barras a fin de visualizar entre que valores se da la mayor acumulación de frecuencias (densidad).

El gráfico de violín representa información numérica en función de variables categóricas, así como los gráficos de caja, pero a diferencia provee de información sobre la densidad de la distribución de los datos, dentro de cada categoría.

R básico tiene funciones para crear un gráfico de densidad, sin embargo, utilizaremos funciones del paquete llamado “ggplot2”, en principio por la estética de la graficación en este paquete. Primeramente vamos a instalar el paquete y a hacerlo disponible:

```
> install.package("ggplot2")  
> library("ggplot2")
```

A continuación, vamos a crear un gráfico de densidad con el uso de los datos de humedad relativa del aire utilizado anteriormente. Para ello haremos uso de la función “ggplot()” más dos argumentos: “data=” con él especificaremos el nombre de la variable donde estan guardados los datos y “aes()” con el que le indicamos al programa que vamos a utilizar los valores de la humedad relativa, guardados en la variable “HR” en el eje X (“x=HR”), y también el tipo de gráfico que deseamos crear, en este caso un gráfico de densidad con la función “geom_density()” (Figura 121 A):

```
> ggplot(data=HR, aes(x=HR)) + geom_density()
```

Notemos que las funciones son un tanto diferentes a las utilizadas anteriormente para la elaboración de los otros gráficos, esto es porque son exclusivas del paquete “ggplot2”. La función “geom_density()” se añade con el signo más (+), ya que el concepto del paquete es el trabajo en capas, con el signo más se añade una capa al gráfico. A continuación vamos a asignar el nombre de los ejes X y Y y un título principal con la función “labs()” (Figura 121 B):

```
> ggplot(data=HR, aes(x=HR)) + geom_density() + labs  
(y="Densidad", x="Humedad Relativa del Aire", title=  
"Distribución de los datos")
```

Aplicaciones de Estadística Básica

En el gráfico se muestra una línea suavizada representando la distribución de los datos, en el cual la mayoría de los valores están concentrados 80 y 90 por ciento de humedad. Si se desea comparar dicha distribución por cada variable categórica “Sitios”, lo podemos hacer fácilmente asignando el argumento “color=Sitios” dentro de la función “aes()”, con el cual básicamente le estamos indicando al programa que divida la distribución por “Sitios” y establezca diferentes colores respectivamente (Figura 121 C):

```
> ggplot(data=HR, aes(x=HR, color=Sitios)) + geom_density() +  
labs(y="Densidad", x="Humedad Relativa del Aire", title=  
"Distribución de los datos")
```

Ahora podemos comparar la distribución de la humedad relativa por sitios. En el Sitio 1 aunque el rango de humedad relativa es amplio (de casi 20 a más de 90%), la densidad de los valores se mantiene constante a lo largo de dicho rango y menor que en los otros sitios; mientras que en el Sitio 2 y Sitio 3 los rangos de valores de la humedad relativa son más cortos y entre 50 a más de 90%, adicionalmente presentan dos picos de mayor densidad de datos, uno pequeño alrededor de 60% y uno más de dos veces mayor, entre 80 y 90%.

No solamente podemos variar el color de los contornos de las curvas de densidad, sino podemos colorear su interior si cambiamos el argumento “color=Sitios” por “fill=Sitios”, además, para presentar un gráfico más estético adicionaremos otros argumentos: Dentro de la función “geom_density()” adicionaremos “col=“gray75”” y “alpha=0.35”, con el primer argumento le indicamos al programa que asigne el color gris 75% a los contornos de las curvas de densidad y con el segundo, le indicamos el nivel de transparencia, en este caso a 0.35 (35%). Finalmente, le indicamos al programa que requerimos la presentación clásica del gráfico, en lugar de la presentación por defecto, esto es cambiar el fondo gris a cuadro por un fondo blanco y las escalas de los ejes en negro con la función “theme_classic()” (Figura 121 D):

```
> ggplot(data=HR, aes(x=HR, fill=Sitios)) + geom_density(  
col="gray75", alpha=0.35) + labs(y="Densidad", x="Humedad  
Relativa del Aire", title="Distribución de los datos") +  
theme_classic()
```

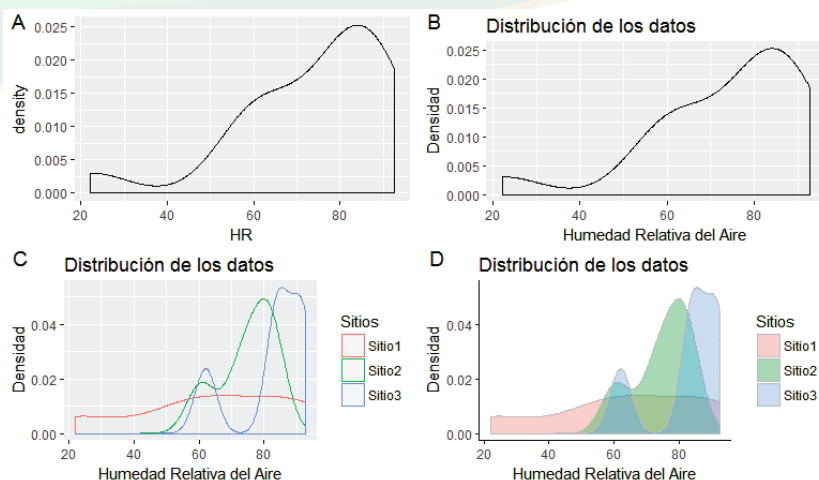


Figura 121. Creación y personalización de gráficos de densidad utilizando datos de humedad relativa del aire (%) y el paquete “ggplot2”. A. Gráfico base; B. Asinación de títulos de los ejes X, Y y título principal; C. Comparación de las distribuciones por sitios, utilizando contornos de colores; D. Mismo que C, pero utilizando relleno de color y tema de gráficos clásicos.

Seguidamente crearemos un gráfico de violín con la función “ggplot()” más dos argumentos: “data=” con él especificaremos el nombre de la variable donde estan guardados los datos y “aes()” con el que le indicamos al programa que vamos a utilizar los valores de la humedad relativa, guardados en la variable “Sitios” en el eje X (“x=Sitios”) y “HR” en el eje Y (“y=HR”), y también le indicamos el tipo de gráfico que deseamos crear en este caso un gráfico de violín con la función “geom_violin()” (Figura 122 A):

```
> ggplot(data=HR, aes(x=Sitios, y=HR)) + geom_violin()
```

El gráfico de violín muestra dos cosas para cada categoría de la variable categórica “Sitios”: 1. El rango de los datos de la variable humedad relativa, y 2. Entre qué valores se concentran las mayores frecuencias de ocurrencia de dichos datos (densidad). En el Sitio 1 la figura es larga y delgada, indicando que aunque el rango de humedad relativa es amplio (de casi 20 a más de 90%) la densidad de los valores se mantiene constante a lo largo de dicho rango y menor que en los otros sitios; mientras que en el Sitio 2 y Sitio 3 los rangos valores de la humedad relativa, son más cortos y entre 60 a más de 90%. Es notorio que para el sitio 2 la mayoría de los valores de humedad relativa se concentran en 80%, mientras que para el sitio 3 la mayor densidad de datos es alrededor de 90%. A gusto personal puedo decir que el gráfico de violín es más preciso que el de densidades e histogramas, en presentar y comparar la distribución de datos.

Aplicaciones de Estadística Básica

Vamos a seguir personalizando el gráfico de violín, añadiendo el nombre de los ejes X y Y y un título principal con la función “labs()” (Figura 122 B):

```
> ggplot(data=HR, aes(x=Sitios, y=HR)) + geom_violin() + labs  
(y="Humedad Relativa del Aire", x="Sitios", title="Humedad  
Relativa por sitios")
```

Seguidamente asignaremos el argumento “color=Sitios” dentro de la función “aes()” con el cual básicamente le estamos indicando al programa que asigne colores diferentes basado en la variable “Sitios” (Figura 122 C):

```
> ggplot(data=HR, aes(x=Sitios, y=HR, color=Sitios)) + geom_  
violin() + labs(y="Humedad Relativa del Aire", x="Sitios",  
title="Humedad Relativa por sitios")
```

Finalmente demostraremos como rellenar las figuras con diferentes colores cambiando el argumento “color=Sitios” por “fill=Sitios”, además, para presentar un gráfico más estético adicionaremos otros argumentos: Dentro de la función “geom_violin()” adicionaremos “col=NA”, con el cual le indicamos al programa que le quite el color de los contornos; además, le indicamos que requerimos la presentación clásica del gráfico en lugar de la presentación por defecto, esto es cambiar el fondo gris con cuadros, por un fondo blanco y las escalas de los ejes en negro, con la función “theme_classic()” (Figura 122 D):

```
> ggplot(data=HR, aes(x=Sitios, y=HR, fill=Sitios)) + geom_violin  
(col=NA) + labs(y="Humedad Relativa del Aire", x="Sitios",  
title="Humedad Relativa por sitios") + theme_classic()
```

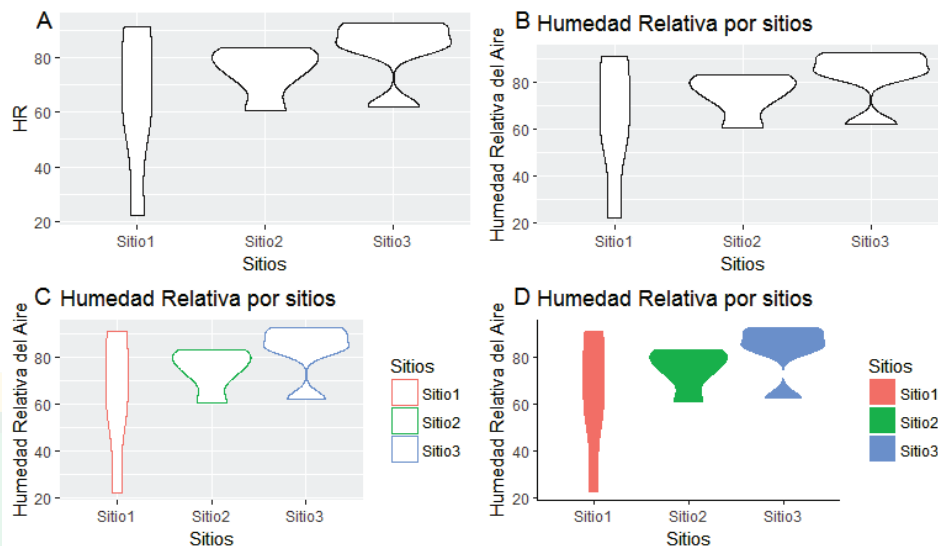


Figura 122. Creación y personalización de gráficos de violín, utilizando datos de humedad relativa del aire (%) y el paquete “ggplot2”. A. Gráfico base; B. Asignación de títulos de los ejes X, Y y título principal; C. Cambio del color de los contornos basado en una variable categórica (Sitios); D. Mismo que C, pero utilizando relleno de color, sin bordes y tema de gráficos clásicos.

Gráfico de doble eje Y

Las relaciones de dos variables numéricas pueden ser representadas con un gráfico de doble eje Y, en cuyo caso los datos se presentan en función de algún tipo de variable categórica o numérica de interés, que por lo general representa al tiempo, pero que no se limita solamente a este. En este gráfico el eje Y de una de las variables se mostrará en el lado izquierdo del gráfico y la otra variable, se mostrará en el lado derecho del gráfico. Para ejemplificar utilizaremos los datos de cobertura del briófito hepático *Porella platyphylla* que se presenta sobre la corteza de los árboles. Pretendemos hacer un gráfico que muestra la cobertura en el eje Y1 y la temperatura del aire en el eje Y2. Los datos los importamos a R y los guardamos en la variable llamada “Porella” (datos en anexo 7):

```
> Porella <- read.csv(file.choose())
> head(Porella)
  Arbol  Sitio Cober Temp    HR   CO2
1     1   Bosque  98.1  10.2  92.5 316.4
2     2   Bosque  86.8  12.4  92.4 313.0
3     3   Bosque  88.8  10.2 104.0 304.9
4     4   Bosque  76.9  15.8  89.0 342.6
5     5   Bosque  83.0  14.5  75.7 322.8
6     6   Bosque  98.1   9.9  91.6 300.2
> unique(Porella$Sitio)
[1] Bosque  Agrícola Urbano
Levels: Agrícola Bosque Urbano
```

En primer lugar, elaboraremos el gráfico de la primera variable (Cober=Cobertura del briófito) (Figura 123 A):

```
> plot(Porella$Cober ~ Porella$Arbol, type="l", xlab="Número
de árboles", ylab="Cobertura (%)", las=1)
```

Luego añadimos el gráfico con la segunda variable (Temp=Temperatura del aire), pero antes utilizaremos el comando “par(new=TRUE)” para indicarle al programa que coloque el segundo gráfico sobre el primero a modo de capa. Para este segundo gráfico, vamos a cambiar el tipo de línea con el argumento “lty=” y vamos a quitar todos los

Aplicaciones de Estadística Básica

títulos y elementos de los ejes con el par de argumentos: “ann=FALSE” y “axes=FALSE” (Figura 123 B):

```
> plot(Porella$Cober ~ Porella$Arbol, type="l", xlab="Número  
de árboles", ylab="Cobertura (%)", las=1)  
> par(new=TRUE)  
> plot(Porella$Temp ~ Porella$Arbol, type="l", lty=2,  
ann=FALSE, axes=FALSE)
```

Seguidamente agregamos los ejes con sus nombres, con el uso de la función “axis()” y dos argumentos, el número “4”, que representa el cuarto lado del cuadro del gráfico, iniciando con el lado de abajo y continuando a favor de las manecillas del reloj; el segundo argumento (las=2) es la posición de los números del eje, en este caso se están estableciendo los números siempre perpendicular al eje. Adicionalmente anexamos el título del eje con la función “mtext()” y tres argumentos: el primero es el título en forma de expresión; el segundo indica el lado y el tercero indica la línea donde se colocará dicho título, en este caso en la línea 3 a partir del borde del gráfico. La codificación completa se muestra a continuación (Figura 123 C):

```
> plot(Porella$Cober ~ Porella$Arbol, type="l", xlab="Número  
de árboles", ylab="Cobertura (%)", las=1)  
> par(new=TRUE)  
> plot(Porella$Temp ~ Porella$Arbol, type="l", lty=2,  
ann=FALSE, axes=FALSE, ylim=c(8,32))  
> axis(4, las=2)  
> mtext(expression("Temperatura" ~degree~C), side=4, padj=3)
```

En caso que el margen derecho no permita la inclusión del título del segundo eje Y, tendríamos que incrementar su tamaño, como sugerencia se podría establecer el comando “par(mar=c(5.1,4.1,4.1,4.1))” al inicio del comando para elaborar el primer gráfico.

Para finalizar el gráfico, se añade la leyenda con la función “legend()” y los argumentos ya conocidos (Figura 123 D):

```
> plot(Porella$Cober ~ Porella$Arbol, type="l", xlab="Número  
de árboles", ylab="Cobertura (%)", las=1)  
> par(new=TRUE)  
> plot(Porella$Temp ~ Porella$Arbol, type="l", lty=2,  
ann=FALSE, axes=FALSE, ylim=c(8,32))  
> axis(4, las=2)  
> mtext(expression("Temperatura" ~degree~C), side=4, padj=3)  
> legend(locator(1), c("Cobertura", "Temperatura"), lty=c(1,2),  
cex=0.5)
```

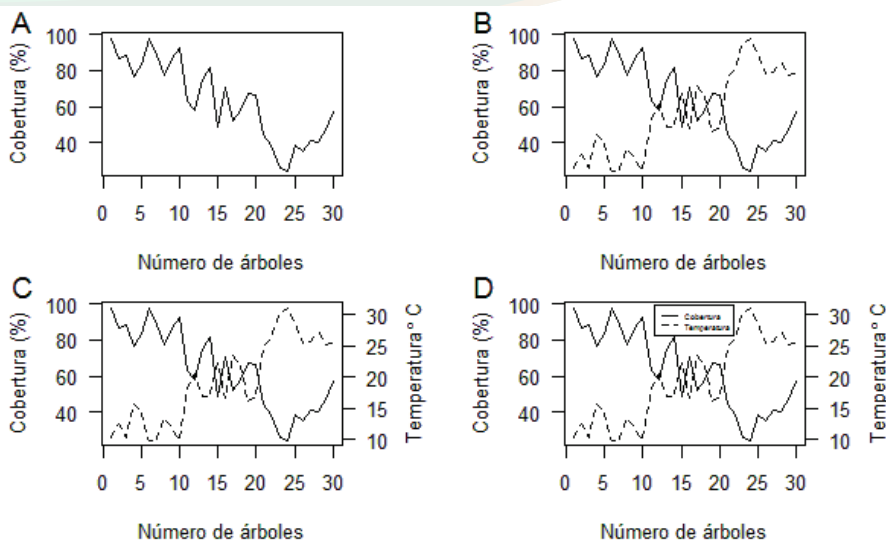



Figura 123. Representación de los pasos para hacer un gráfico con doble eje Y. A. Elaboración de gráfico de la primera variable; B. Inserción del gráfico de la segunda variable; C. Asignación de ejes, nombres y título de ejes; D. Asignación de la leyenda.

Gráfico multipaneles

Representar varios gráficos en una sola figura es esencial en toda publicación, pues ahorra espacio y ayuda a determinar relaciones e interacciones visualmente. En R básico, vamos a crear paneles con el argumento “mfrow=” o “mfcoll” dentro de la función “par()”. También utilizaremos la función “layout()”. Es bueno reconocer que para elaborar paneles de una forma más fácil y rápida se pueden utilizar paquetes con opciones gráficas que facilitan su elaboración, incluidos lattice y ggplot2.

En principio vamos a preparar cuatro gráficos que serán utilizados para crear un panel de cuatro graficos, estos gráfico son: un gráfico de pastel, un histograma, un gráfico de barras y un gráfico de líneas, todos con diferentes datos. A continuación se explica la preparación de cada uno de los gráficos.

1. El gráfico de pastel lo prepararemos con una lista de 120 animales clasificados a nivel de clases (Anexo 5). Los datos los guardamos primeramente en una variable llamada “Datos1” y luego hacemos la cuenta de las frecuencias de observación en cada clase, utilizando la función “table()” y el resultado lo guardamos en la variable “Pastel”, quedando de esa forma lista la variable para ser utilizada como primer grafico:

Aplicaciones de Estadística Básica

```
> Datos1 <-read.csv(file.choose())
> head(Datos1)
      Clases
1 Mamífero
2 Mamífero
3 Mamífero
4 Mamífero
5 Mamífero
6 Mamífero
> Pastel <-table(Datos1)
> Pastel
Datos1
Anfibio      Ave      Mamífero      Reptil
      6      88      12      14
```

2. El histograma lo elaboraremos con los datos de altura de 10 personas medidas en centímetros. Los datos primeramente los almacenaremos en la variable “Datos2”, luego extraeremos solamente la columna de valores de altura y la guardaremos en la variable “Histograma”, quedando de esa forma lista la variable para ser utilizada como segundo grafico:

```
> Datos2 <-read.csv(file.choose())
> head(Datos2)
  Personas Altura
1 Persona1    143
2 Persona2    153
3 Persona3    160
4 Persona4    162
5 Persona5    165
6 Persona6    168
> Histograma <-Datos2[,2]
> Histograma
[1] 143 153 160 162 165 168 172 173 177 180
```

3. El gráfico de barra lo construiremos con datos de Humedad Relativa del Aire en porcentajes (HR) (arreglados por filas), la cual guardaremos primeramente en la variable llamada “Datos3”, seguidamente se calcularán las medias en función de los sitios utilizando la función “tapply()” y los resultados los guardaremos en la variable “Graf_Bar”, quedando de esa forma lista la variable para ser utilizada como tercer grafico:

```
> Datos3 <-read.csv(file.choose())
> head(Datos3)
  Sitios   HR
1 Sitio1 62.3
2 Sitio1 91.2
3 Sitio1 56.9
4 Sitio1 84.3
5 Sitio1 22.1
6 Sitio2 81.2
> Graf_Bar <-tapply(Datos3$HR, Datos3$Sitios, mean)
> Graf_Bar
Sitio1 Sitio2 Sitio3
 63.36  74.92  82.72
```

4. Para el gráfico de línea utilizaremos los datos de Oxígeno Disuelto en 10 sitios medidos por varios meses. Para este ejemplo, solo utilizaremos el mes de abril y los sitios, por lo que guardamos solamente la columna 1 y 2 en la variable “Graf_Linea”, quedando de esa forma lista la variable para ser utilizada como cuarto grafico:

```
> Datos4 <-read.csv(file.choose())
> head(Datos4)
  Sitios Abr May Jun Jul Ago
1      1 5.6 4.5 4.7 5.7 7.0
2      2 4.2 5.5 6.1 6.5 7.5
3      3 3.1 4.6 4.6 5.4 7.7
4      4 3.3 4.6 4.8 5.8 7.9
5      5 6.3 5.4 5.6 6.4 7.2
6      6 3.5 5.2 6.0 5.8 8.1
7      7 2.3 3.6 3.8 4.6 6.9
8      8 3.2 3.2 4.8 6.7 6.8
9      9 4.1 5.4 5.6 6.4 7.3
10     10 8.2 2.3 3.9 5.5 6.7
> Graf_Linea <-Datos4[,1:2]
> head(Graf_Linea)
  Sitios Abr
1      1 5.6
2      2 4.2
3      3 3.1
4      4 3.3
5      5 6.3
6      6 3.5
```

Aplicaciones de Estadística Básica

Después de preparar los gráficos que se utilizarán como paneles para el gráfico de multipaneles, procederemos a indicarle al programa que divida la pantalla del gráfico en cuatro regiones, utilizando la función “par()” y el argumento “mfrow=” o “mfcoll=”, la diferencia entre estos dos últimos argumentos estriba en que con “mfrow=” el programa muestra los gráficos ordenados por fila; por otro lado, con “mfcoll=” el programa muestra los gráficos ordenados por columnas (Figura 124). El número de columnas y filas en que se dividirá la pantalla de gráfico, lo indicaremos al programa con un vector de dos números, el primero representa el número de filas y el segundo el número de columnas. En el caso del ejemplo pretendemos formar cuatro paneles ordenados en dos filas y dos columnas, de tal forma que el vector lo escribiremos como “c(2,2)”.

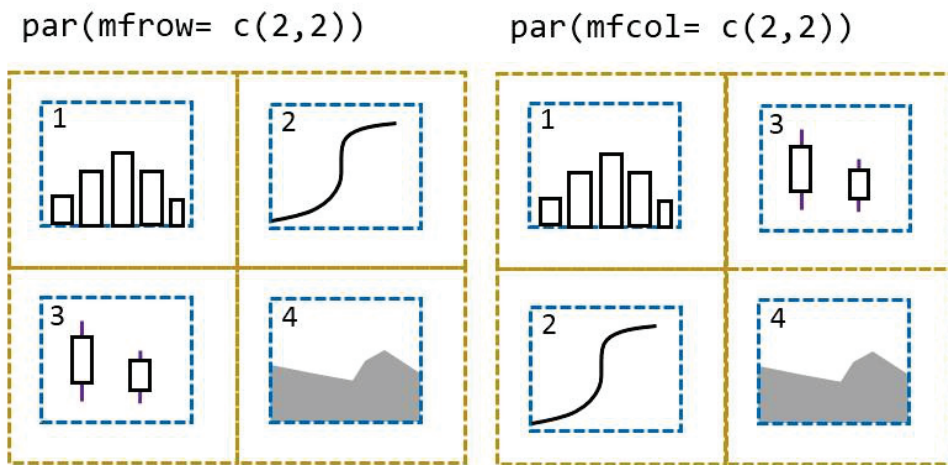


Figura 124. Ilustración del como los argumentos “mfrow=” y “mfcoll=” insertan los gráficos de diferente forma. El argumento “mfrow=” inserta los gráficos siguiendo un orden por filas; el argumento “mfcoll=” inserta los gráficos siguiendo un orden por columnas.

Con todo lo antes preparado y conocido, se puede elaborar el gráfico de multipaneles mediante cinco líneas de comando. Ejemplificaremos el caso de un panel arreglado por filas (Figura 125 A):

```
> par(mfrow=c(2,2))
> pie(Pastel)
> hist(Histograma)
> barplot(Graf_Bar)
> plot(Graf_Linea$Abr ~ Graf_Linea$Sitios, type="l")
```

Ejemplificaremos el caso de un panel arreglado por columnas (Figura 125 B):

```
> par(mfcol=c(2,2))
> pie(Pastel)
> hist(Histograma)
> barplot(Graf_Bar)
> plot(Graf_Linea$Abr ~ Graf_Linea$Sitios, type="l")
```

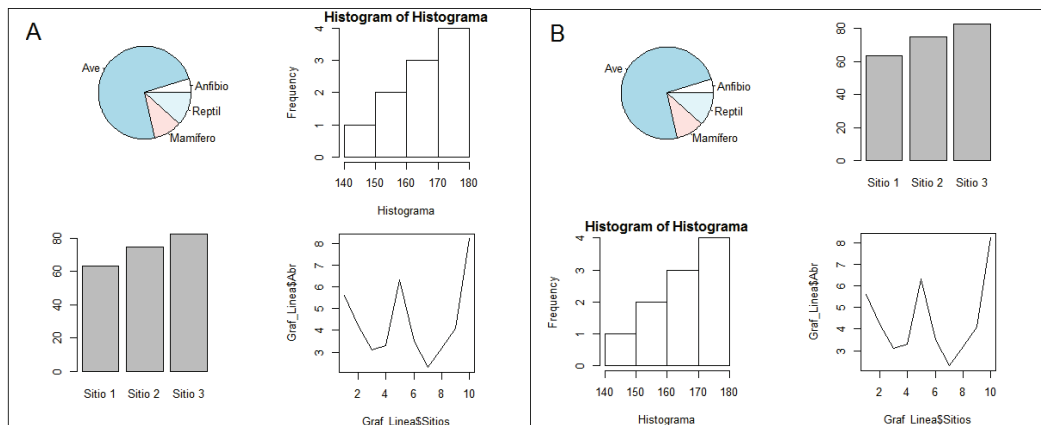


Figura 125. Gráficos multipanel utilizando los argumentos “mfrow=” o “mfcol=” dentro de la función “par()”. A. Ordenados por filas (mfrow=c(2,2)) y B. Ordenado por columnas (mfcol=c(2,2)).

Otra forma de crear paneles un poco más creativos, es con el uso de la función “layout()”, con esta podemos combinar los espacios de los paneles para crearlos más grandes. Para utilizar la función tenemos que crear una tabla que le indique al programa la forma en que se arreglarán los paneles. La tabla se estructura con la función “rbind()” y sus argumentos son dos vectores que definen los arreglos de los datos. Por ejemplo, si estamos interesados en que el gráfico de multipaneles presente un solo panel en la parte superior y dos en la parte inferior, los vectores a combinar serían “c(1,1)” y “c(2,3)”, así estructuramos una tabla donde el gráfico 1 se mostrará en la primera fila, y el dos y el tres en la segunda fila (Figura 126). La tabla la guardamos en una variable a la que llamaremos “Presentación1”:

```
> Presentación1 <-rbind(c(1,1),c(2,3))
> Presentación1
      [,1] [,2]
[1,]    1    1
[2,]    2    3
```

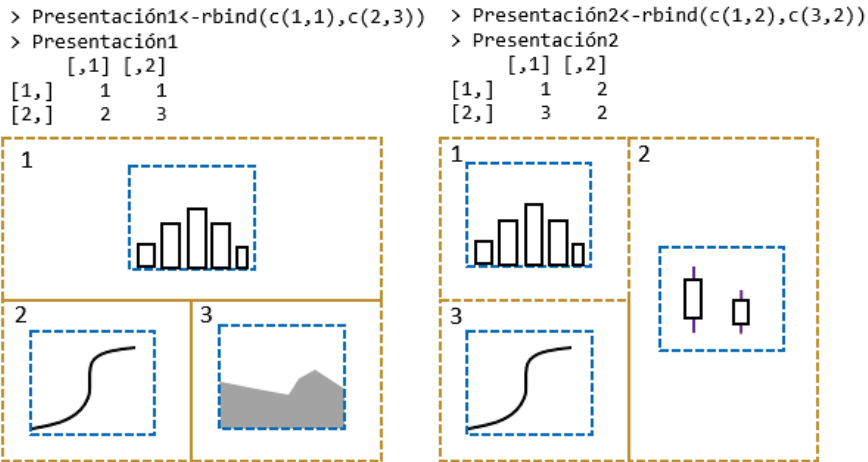


Figura 126. Ejemplo del uso de la función “layout()” combinando paneles. A. Combinación de los paneles en la primera fila; B. Combinación de los paneles al lado derecho.

Una vez construidas las tablas a utilizar con la función “layout()”, procedemos a asignar los gráficos, el primer gráfico asignado ocupará el espacio 1 designado en la tabla, el segundo gráfico asignado ocupará el espacio 2 y así sucesivamente (Figura 127 A):

```
> layout(Presentación1)
> hist(Histograma)
> barplot(Graf_Bar)
> plot(Graf_Linea$Abr ~ Graf_Linea$Meses, type="l")
```

Si se desea que el panel derecho esté combinado para alojar a un único gráfico, tenemos que arreglar la tabla de tal forma que el programa entienda y despliegue lo que nosotros queremos, por ejemplo:

```
> Presentación2 <-rbind(c(1,2),c(3,2))
> Presentación2
      [,1] [,2]
[1,]    1    2
[2,]    3    2
```

De igual forma utilizamos la función “layout()” con la tabla guardada en la variable “Presentación2” (Figura 127 B):

```
> layout(Presentación2)
> hist(Histograma)
> barplot(Graf_Bar)
> plot(Graf_Linea$Abr ~ Graf_Linea$Meses, type="l")
```

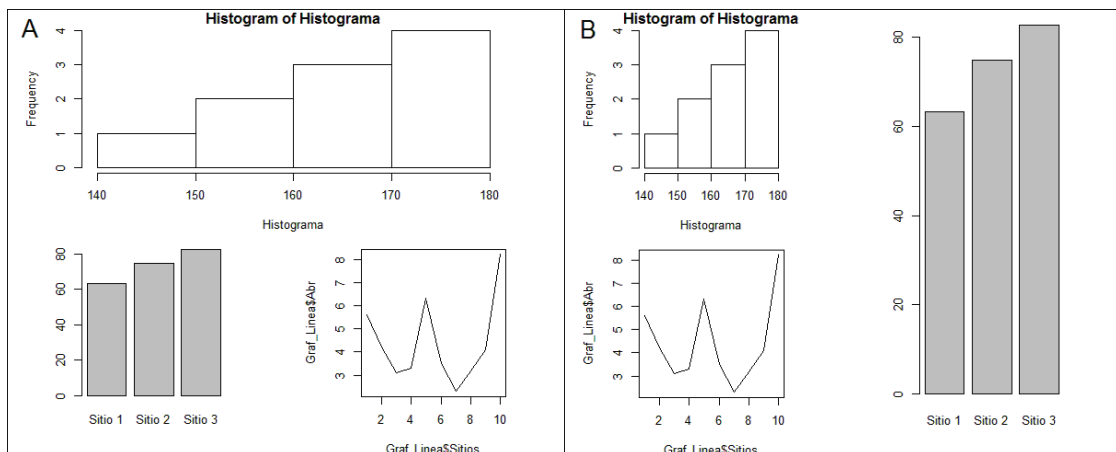


Figura 127. Gráficos multipanel utilizando la función “`layout()`”. A. Paneles superiores combinados y B. paneles derechos combinados.

En los paneles combinados los gráficos aparecen estirados para ajustarlos al tamaño de dichos paneles, sin embargo, existen unos comandos que permiten evitar que los gráficos se estiren, y que por el contrario mantengan su forma cuadrada. Para el segundo ejemplo (Figura 127 B), con la función “`par()`” y el argumento “`pty="s"`”, o sea `par(pty="s")` antes de la línea “`barplot(Graf_Bar)`” podemos mantener cuadrado el gráfico de barra, tomando en cuenta que antes que se aplique la línea del siguiente comando, o sea “`plot(Graf_Linea$Abr ~ Graf_Linea$Meses, type="l")`”, el argumento debe regresarse a `par(pty="m")`, que es la forma por defecto. Si se desea regresar a visualizar, solamente un gráfico, es necesario regresar a la opción por defecto con el comando “`layout(1)`”.

Cuando el número de paneles aumenta dentro del gráfico, la personalización de los elementos de los gráficos con las funciones básicas de R se hace muy cargada en términos de la cantidad de comando, es aquí donde se sugiere el uso del algún paquete que facilite este trabajo. Para demostrarlo utilizaremos los paquetes “`lattice`” y “`ggplot2`”. Los datos a emplear son los de oxígeno disuelto en 10 sitios por cinco meses, los cuales están arreglados por filas (datos completos en anexo 8) y que en principio los guardaremos en la variable “`OD_Meses`”:

Aplicaciones de Estadística Básica

```
> OD_Meses <-read.csv(file.choose())
> head(OD_Meses)
  Sitios Meses  OD
1      1 Abril 5.6
2      2 Abril 4.2
3      3 Abril 3.1
4      4 Abril 3.3
5      5 Abril 6.3
6      6 Abril 3.5
> unique(OD_Meses$Meses)
[1] Abril  Mayo  Junio Julio  Agosto
Levels: Abril Agosto Julio Junio Mayo
```

Para utilizar “lattice”, primeramente instalemos el paquete mediante la función “install.packages()” y lo hacemos disponible con la función “library()”:

```
> install.packages("lattice")
> library("lattice")
```

El gráfico multipanel lo creamos con una sola línea de comando, utilizando la función “xyplot()” y tres argumentos: la primera parte del primer argumento define los datos a usar (los valores OD en función de los sitios), la segunda parte del primer argumento separado por “|” le indica que el factor (variable categórica con la que se van a crear los paneles) es la columna “Meses” (abril, mayo, junio, julio, agosto); el segundo argumento indica el tipo de gráfico, en este caso “l” representa gráfico de línea; y el tercer argumento le indica la variable donde está almacenada la información (Figura 128 A):

```
> xyplot(OD ~ Sitios | factor(Meses), type="l", data=OD_Meses)
```

Inmediatamente nos damos cuenta que el panel que se creó por defecto presenta dos inconvenientes, el primero es que tiene dos paneles arriba y tres abajo, el segundo es que los meses no están ordenadamente, aunque dentro de la base de datos aparezcan ordenados. Para que nos presente tres paneles en la parte superior y dos en la parte inferior, utilizaremos el argumento “as.table=TRUE” (Figura 128 B):

```
> xyplot(OD ~ Sitios | factor(Meses), type="l", data=OD_Meses,
as.table=TRUE)
```

Seguidamente diseñaremos un vector con el orden de los meses con la función “paste0()” y lo guardamos en la variable que llamaremos “MesesO”:

```
MesesO <-paste0(c("Abril", "Mayo", "Junio", "Julio", "Agosto"))
```


Luego integramos el vector elaborado y almacenado en la variable “MesesO” en la función “factor()” como un nuevo argumento llamado “levels=” (Figura 128 C):

```
> xyplot(OD ~ Sitios | factor(Meses, levels=MesesO), type="l",  
data=OD_Meses, as.table=TRUE)
```

Notemos que los paneles por meses, ahora si están ordenados de forma correcta. Adicionalmente se puede cambiar el color de las líneas y de la franja de los nombres con los argumentos “col.line=” y “strip=”, para este último argumento se establece la función “strip.custom()”, dentro de la cual se coloca otro argumento llamado “bg=” que es una abreviatura que significa fondo (background o fondo en español) y a esta se le asigna el color (Figura 128 D):

```
> xyplot(OD ~ Sitios | factor(Meses, levels=MesesO), type="l", data=OD_  
Meses, as.table=TRUE, col.line="red", strip=strip.custom(bg="green"))
```

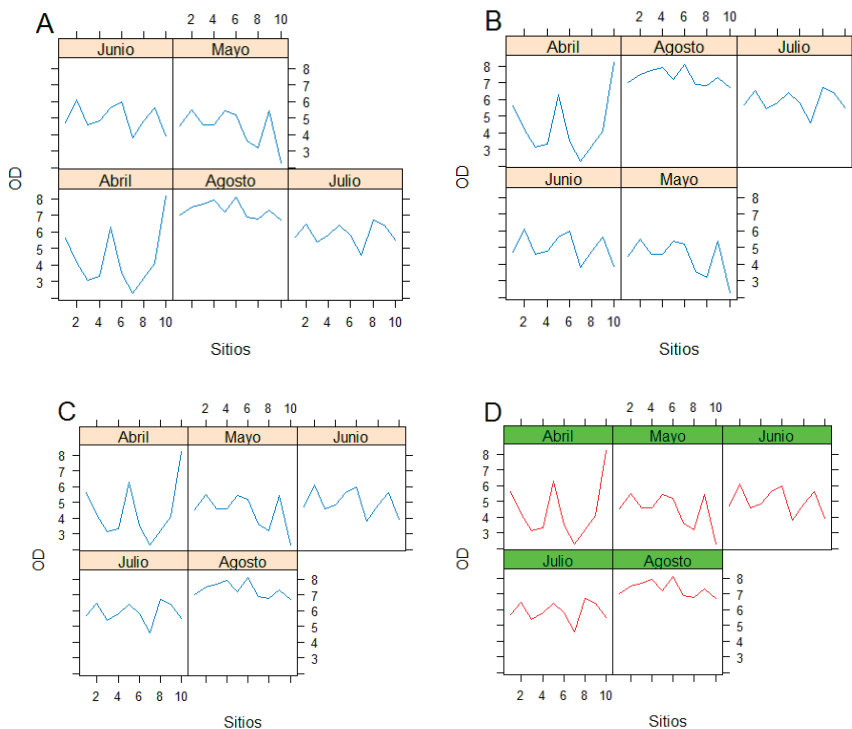


Figura 128. Gráficos multipaneles elaborados con el uso de la función “xyplot()” del paquete “lattice”. A. Gráfico por defecto; B. Gráfico con paneles ordenados de tal forma que haya tres arriba y dos abajo; C. Paneles ordenados por meses; D. Personalización del color de las líneas y de las bandas de los nombres.

Aplicaciones de Estadística Básica

Los paneles en el gráfico los podemos modificar fácilmente, utilizando el argumento “layout=c()” y especificando el número de columnas y filas. En este caso, el primer valor equivale al número de columnas y el segundo al número de filas, de tal forma que si queremos que los paneles se arreglen en cinco columnas y una fila, el argumento completo quedaría escrito como “layout=c(5,1)” (Figura 129 A):

```
> xyplot(OD ~ Sitios | factor(Meses, levels=Meses0), type="l",  
layout=c(5,1), data=OD_Meses)
```

Y si queremos que los paneles se arreglen en una columna y cinco filas, el argumento completo quedaría escrito como “layout=c(1,5)” (Figura 129 B):

```
> xyplot(OD ~ Sitios | factor(Meses, levels=Meses0), type="l",  
layout=c(1,5), data=OD_Meses)
```

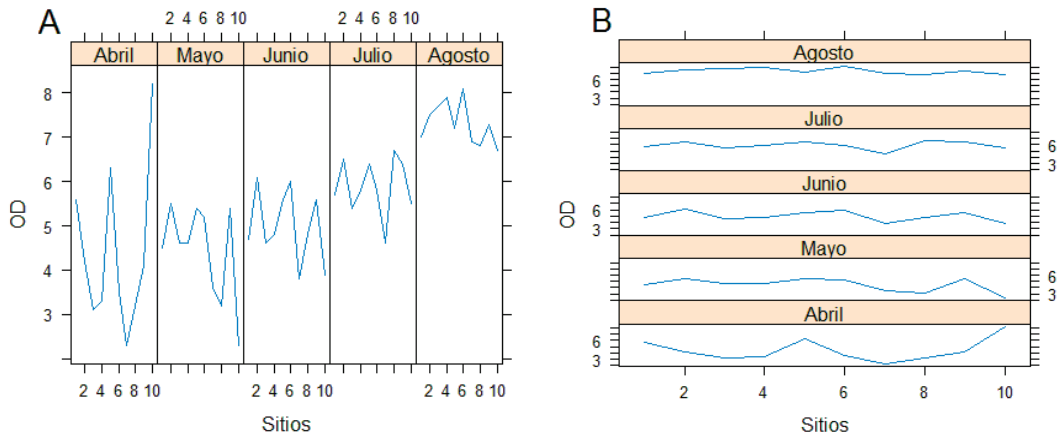


Figura 129. Otros ejemplos de gráficos multipaneles elaborados con el uso de la función “xyplot()” del paquete “lattice”. A. Incluyendo el argumento “layout=c(5,1)”; B. Incluyendo el argumento “layout=c(1,5)”.

Para utilizar “ggplot2”, primeramente instalemos el paquete mediante la función “install.packages()” y lo hacemos disponible con la función “library()”:

```
> install.packages("ggplot2")  
> library("ggplot2")
```

El paquete “ggplot2” es específico para crear gráficos de excelente calidad, utilizando una exclusiva gramática. Los gráficos se elaboran por capas, en las cuales el usuario va añadiendo de forma personalizada. Antes de demostrar cómo se crean gráficos multi-

paneles, vamos a explorar la información de nuestros datos de Oxígeno Disuelto, con un gráfico normal de líneas, para esto utilizamos la función “ggplot()”, en la cual colocamos varios argumentos: la variable donde se encuentran los datos (OD_Meses) y la función “aes()” (aesthetic o estética en español).

En la función “aes()” le indicamos al programa la información a utilizar. En el eje X asignamos a la variable “Sitios” como factor (variable categórica) escribiendo “x=factor(Sitios)”; en el eje Y asignamos a la variable “OD” como datos cuantitativos escribiendo “y=OD”; y le decimos que agrupe los datos con la variable “Meses” escribiendo “group=Meses”. Todas estas asignaciones corresponden con la primera capa del gráfico. La segunda capa la añadimos con el signo más (+) y el argumento “geom_line()” que le indica al programa que dibuje líneas; en la tercera capa le agregamos las etiquetas utilizando la función “labs()” y asignando los nombres de los ejes X y Y. El comando completo quedaría escrito de la siguiente forma (Figura 130 A):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses)) +  
geom_line() + labs(x="Sitios de muestreo", y="Oxígeno  
Disuelto")
```

Si quisiéramos hacer el gráfico un poco más colorido, podemos utilizar el argumento “col=” dentro de la función “aes()” y en lugar de asignar los colores, le asignamos la variable “Meses” para que el programa asigne un color diferentes (automáticamente) por cada uno de los meses (Figura 130 B):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses, col=Meses))  
+ geom_line() + labs(y="Oxígeno Disuelto", x="Sitios de muestreo")
```

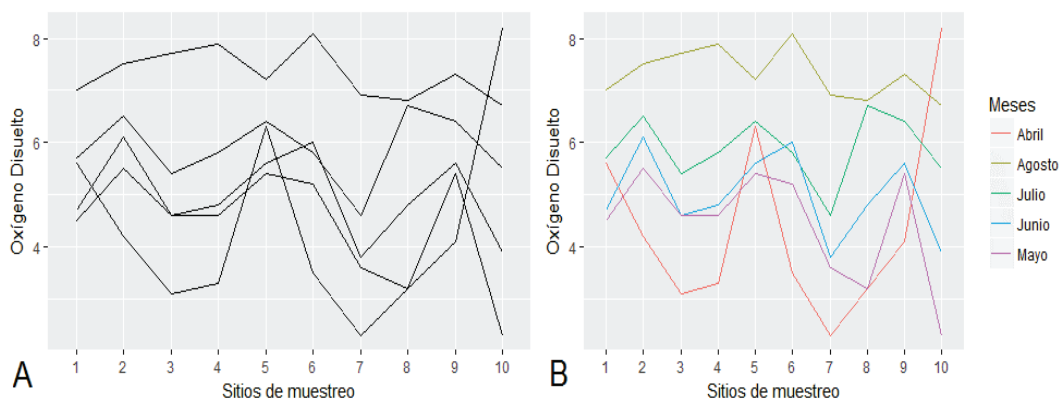


Figura 130. Gráfico de múltiples líneas. A. Gráfico exploratorio con características por defecto; B. Asignación de colores basados en la variable “Meses”.

Aplicaciones de Estadística Básica

Notemos que el orden de los meses en la leyenda no es el que esperamos (abril, mayo, junio, julio, agosto), para restablecer el orden deseado habrá que insertar un factor dentro de los datos de la columna “Meses” con el cual le indicamos al programa el orden deseado, esto será explicado adelante.

Seguidamente anexaremos la función “`facet_wrap()`” para elaborar un gráfico multipanel en el que cada panel corresponderá con la información de Oxígeno Disuelto por meses. La función quedaría escrita como “`facet_wrap(~ Meses)`” en donde “`~ Meses`” quiere decir: “En función de la variable Meses”. El comando completo quedaría escrito de la siguiente manera (Figura 131 A):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses,
col=Meses)) + geom_line() + labs(y="Oxígeno Disuelto",
x="Sitios de muestreo") + facet_wrap(~ Meses)
```

De la misma forma que con el uso del paquete “`lattice`”, notamos rápidamente que los paneles no están arreglados en orden de los meses, a como estaba establecido en la base de datos. De tal forma que tenemos que reorganizar la secuencia de los paneles asignando un factor a la variable “Meses” y añadiendo el argumento “`levels=`” con el cual asignamos el orden, posteriormente volvemos a ejecutar el comando para elaborar el gráfico (Figura 131 B). Con el factor creado, también se pueden reordenar los meses en la leyenda de la figura 130 B.

```
> OD_Meses$Meses <-factor(OD_Meses$Meses, levels=c("Abril", "Mayo",
"Junio", "Julio", "Agosto"))
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses,
col=Meses)) + geom_line() + labs(y="Oxígeno Disuelto",
x="Sitios de muestreo") + facet_wrap(~ Meses)
```

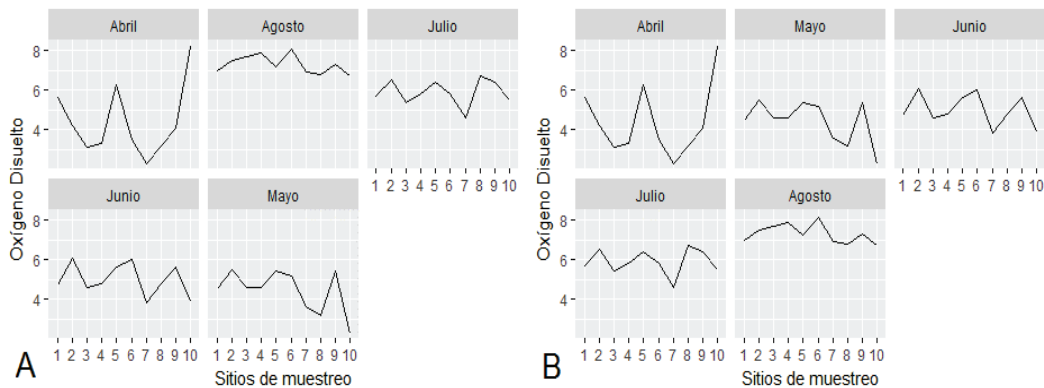


Figura 131. Gráfico multipaneles en `ggplot2`. A. Orden de los gráficos por defecto; B. Orden personalizado.

Con la función “`facet_wrap()`” podemos personalizar no solo el orden, sino la disposición de los paneles utilizando el argumento “`ncol=`” y asignando el número de columnas, donde deseamos que aparezcan los paneles o el argumento “`nrow=`” para asignar el número de filas, donde queremos que aparezcan los paneles. Ejemplificaremos primero para personalizar el número de columnas, haciendo un gráfico multipanel de 4 columnas (Figura 132 A):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses, col=Meses)) + geom_line() + labs(y="Oxígeno Disuelto", x="Sitios de muestreo") + facet_wrap(~ Meses, ncol=4)
```

Seguidamente ejemplificaremos la personalización, según el número de filas, haciendo un gráfico multipanel de 3 filas (Figura 132 B):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses, col=Meses)) + geom_line() + labs(y="Oxígeno Disuelto", x="Sitios de muestreo") + facet_wrap(~ Meses, nrow=3)
```

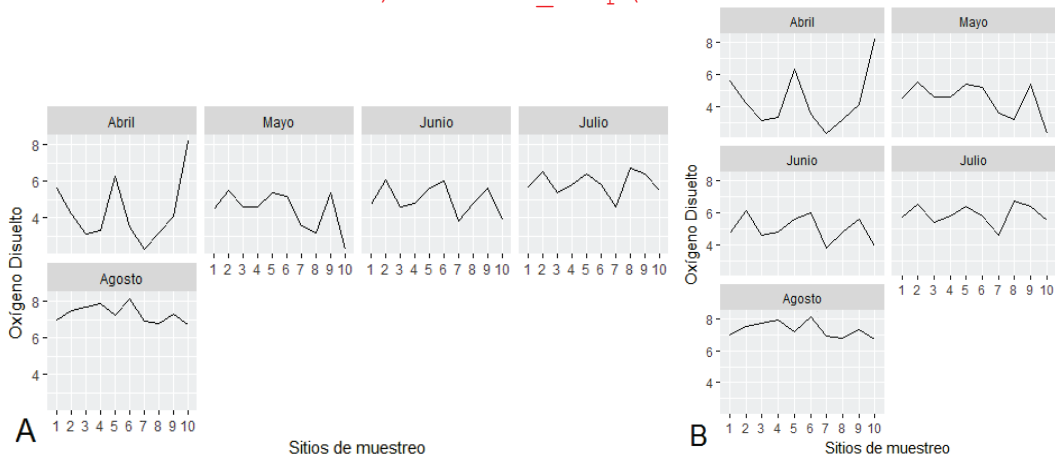


Figura 132. Cambio en la disposición de los paneles dentro del gráfico. A. Arreglado en cuatro columnas; B. Arreglado en tres filas.

Además de la función “`facet_wrap()`”, el paquete `ggplot2` también cuenta con la función “`facet_grid()`” que genera un gráfico multipaneles con visualizaciones por filas o columnas. En el comando quitaremos el argumento “`facet_wrap(~ Meses)`” y lo sustituiremos por “`facet_grid(~ Meses)`”. El punto y la tilde (~) dentro del argumento es lo que determina si el arreglo lo realizamos por columnas o por filas, de tal forma que si lo escribimos “`~ Meses`” los paneles aparecerán arreglados por columnas y si lo escribimos “`Meses~.`” los paneles aparecerán arreglados por fila.

Aplicaciones de Estadística Básica

El comando para la creación de un gráfico multipanel arreglado por columnas, con la función “`facet_grid()`”, quedaría escrito de la siguiente forma (Figura 133 A):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses,
col=Meses)) + geom_line() + labs(y="Oxígeno Disuelto",
x="Sitios de muestreo") + facet_grid(.~ Meses)
```

El comando para la creación de un gráfico multipanel arreglado por filas con la función “`facet_grid()`”, quedaría escrito de la siguiente forma (Figura 133 B):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses,
col=Meses)) + geom_line() + labs(y="Oxígeno Disuelto",
x="Sitios de muestreo") + facet_grid(Meses ~.)
```

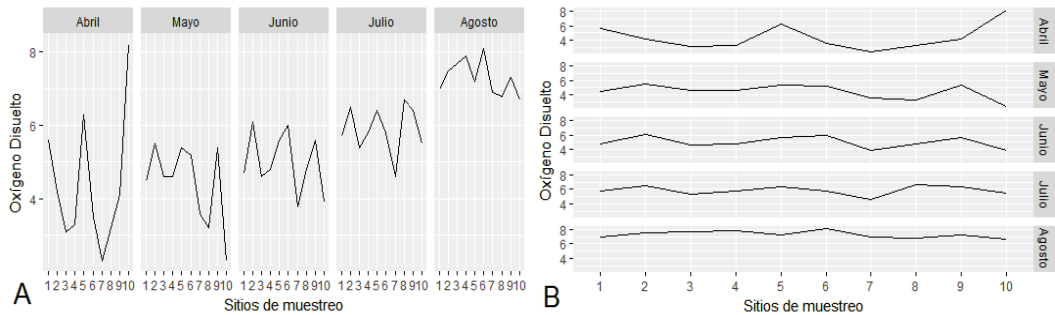


Figura 133. Gráfico multipanel utilizando la función “`facet_grid()`”. A. Arreglo en columnas; B. Arreglo en filas.

Regresando al gráfico multipanel de la figura 131 B, podemos hacer una última personalización para dejar un gráfico en blanco y negro listo para una publicación científica. Para ello vamos a quitar el fondo de los paneles, a asignar las escalas de los ejes y a quitar el borde de la caja de los nombres de los meses de cada panel. Con tal fin, incluimos en el comando las funciones “`theme_classic()`” y “`theme(strip.background=element_rect(colour="white"))`”, la primera cambia el tema que tiene por defecto el gráfico a uno, que es el clásico en blanco y negro para publicaciones; el segundo torna blanco el borde de las cajas que encierran a los títulos de cada panel, si no se incluye esta función con estos argumentos, los nombres estarán rodeado de una caja cuadrada de borde negro, que le restará estética al gráfico multipanel. La codificación completa quedaría expresada de la siguiente manera (Figura 134):

```
> ggplot(OD_Meses, aes(x=factor(Sitios), y=OD, group=Meses)) +
geom_line() + labs(y="Oxígeno Disuelto", x="Sitios de muestreo") +
facet_wrap(~ Meses) + theme_classic() + theme(strip.background=
element_rect(colour="white"))
```

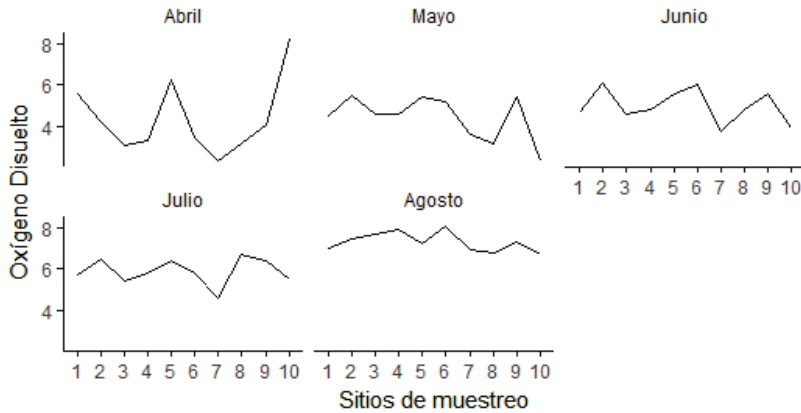


Figura 134. Gráfico multipanel acabado y listo para publicación.

La personalización avanzada de los gráficos utilizando los paquetes “lattice” y “ggplot2” no será un tópico a tratar a profundidad en este escrito. Dado que cada paquete tiene sus propias funciones, argumentos y forma de escribir los comandos, queda en manos del lector, explorar el uso de ellos.

Referencias

Carlberg, C. (2011). Statistical Analysis: Microsoft® Excel 2010. Pearson Education, Inc. Indianapolis, USA.

Dalgaard, P. (2002). Introductory Statistics with R. Springer-Verlag New York, Inc. USA.
Garmendia, M. (2018). Bases de Datos en Microsoft Excel, Diseño y Administración. Universidad Nacional Agraria, Managua, Nicaragua.

Kiernan, D. (2010). Introductory Statistics for Environmental Sciences. Lecture Supplement and Workbook. Second Print. Kendall Hunt Publishing Company. USA.

Kiernan, D. & Bevilacqua, E. (2011). Graduate Student. Statistical Analysis. Handbook. Kendall Hunt Publishing Company. USA.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang & C. Lin. (2017). Package 'e1071'. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

R Core Team (2018a). R: A language and environment for statistical computing. Getting Help with R. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.r-project.org/help.html>

R Core Team (2018b). R: A language and environment for statistical computing. The R manuals. R Foundation for Statistical Computing, Vienna, Austria. URL <https://cran.r-project.org/manuals.html>

Seefeld, K. and E. Linder. (2007). Statistics Using R with Biological Examples. Department of Mathematics & Statistics. University of New Hampshire, Durham, NH, USA.

Teetor, P. (2011). R Cookbook, Proven Recipes for Data Analysis, Statistics, and Graphics. O'Reilly. USA.

Wickham, H. (2016). ggplot2. Elegant Graphics for Data Analysis. Second Edition. Springer. USA.

Zuur, A., E. Ieno, E. Meesters. (2009). A Beginner's Guide to R. Springer Science+Media, LLC. New York. USA.

Anexos

Anexo 1. Datos sobre humedad de suelo (HS) %, tomados en 10 puntos de muestreos en tres diferentes sitios.

Nº	Sitios	HS	Nº	Sitios	HS	Nº	Sitios	HS	Nº	Sitios	HS
1	Sitio1	85.4	9	Sitio1	96.1	17	Sitio2	76.2	25	Sitio3	85.8
2	Sitio1	91.2	10	Sitio1	83.2	18	Sitio2	71.2	26	Sitio3	97.5
3	Sitio1	93.4	11	Sitio2	81.2	19	Sitio2	81.2	27	Sitio3	93.6
4	Sitio1	84.3	12	Sitio2	72.1	20	Sitio2	82.3	28	Sitio3	76.5
5	Sitio1	86.5	13	Sitio2	82.3	21	Sitio3	84.7	29	Sitio3	95.4
6	Sitio1	98.2	14	Sitio2	83.4	22	Sitio3	90.5	30	Sitio3	82.5
7	Sitio1	94.3	15	Sitio2	90.1	23	Sitio3	92.7			
8	Sitio1	77.2	16	Sitio2	81.3	24	Sitio3	83.6			

Anexo 2. Datos de peso (libras) de un grupo de 10 venados, tomados durante los tres momentos.

Nº	ID	Momentos	Peso	Nº	ID	Momentos	Peso
1	1	Momento1	78.5	16	6	Momento2	110.5
2	2	Momento1	123.2	17	7	Momento2	89.4
3	3	Momento1	150.8	18	8	Momento2	109.4
4	4	Momento1	121.3	19	9	Momento2	151.9
5	5	Momento1	79.8	20	10	Momento2	97.7
6	6	Momento1	98.5	21	1	Momento3	89.7
7	7	Momento1	89.3	22	2	Momento3	142.3
8	8	Momento1	102.5	23	3	Momento3	161.2
9	9	Momento1	145.6	24	4	Momento3	123.8
10	10	Momento1	89.9	25	5	Momento3	89.6
11	1	Momento2	84.2	26	6	Momento3	128.7
12	2	Momento2	130.4	27	7	Momento3	105.3
13	3	Momento2	151.3	28	8	Momento3	110.4
14	4	Momento2	122.3	29	9	Momento3	162.3
15	5	Momento2	86.7	30	10	Momento3	109.1

Anexo 3. Datos de concentración de Oxígeno disuelto (OD) (ppm) en el agua, a lo largo de un río principal en una microcuenca. Los muestreos se realizaron en las tres partes

Aplicaciones de Estadística Básica

de la microcuenca: parte alta, parte media y parte baja; adicionalmente, en cada parte se seleccionan dos usos de suelo, estos son el uso bosque y el uso agrícola, y en cada uno de estos usos, se establecieron cinco puntos de muestreo (réplicas).

Nº	Toposec	Ecosist	OD	Nº	Toposec	Ecosist	OD
1	Alta	Bosque	4.5	16	Media	Agrícola	3.5
2	Alta	Bosque	5.3	17	Media	Agrícola	2.3
3	Alta	Bosque	4.3	18	Media	Agrícola	1.5
4	Alta	Bosque	4.8	19	Media	Agrícola	2.6
5	Alta	Bosque	4.7	20	Media	Agrícola	1.3
6	Alta	Agrícola	3.7	21	Baja	Bosque	3.1
7	Alta	Agrícola	4.5	22	Baja	Bosque	4
8	Alta	Agrícola	3.5	23	Baja	Bosque	2.3
9	Alta	Agrícola	4	24	Baja	Bosque	2.4
10	Alta	Agrícola	3.9	25	Baja	Bosque	3.1
11	Media	Bosque	4.3	26	Baja	Agrícola	2.3
12	Media	Bosque	3.1	27	Baja	Agrícola	3.2
13	Media	Bosque	2.3	28	Baja	Agrícola	1.5
14	Media	Bosque	3.4	29	Baja	Agrícola	1.6
15	Media	Bosque	2.1	30	Baja	Agrícola	2.3

Anexo 4. Datos de Oxígeno Disuelto (OD) tomado en la orilla de diferentes ríos, en las tres partes de una microcuenca: parte alta, parte media y parte baja; se realizó un muestreo en cada uno de cinco ríos, en cada parte de la microcuenca.

Nº	Toposec	Ríos	OD
1	Alta	Río Grande	5.6
2	Alta	Río Escondido	4.5
3	Alta	Río El Salto	4.2
4	Alta	Río Alegre	5.3
5	Alta	Río San Luis	2.1
6	Media	Río Grande	4.3
7	Media	Río Escondido	4.2
8	Media	Río El Salto	3.8
9	Media	Río Alegre	4.2
10	Media	Río San Luis	3.4
11	Baja	Río Grande	3.2
12	Baja	Río Escondido	3.6
13	Baja	Río El Salto	3.2
14	Baja	Río Alegre	4.1
15	Baja	Río San Luis	2.9

Anexo 5. Lista de clases taxonómicas, a las cuales pertenecen 120 especies, las clases son mamíferos, aves, reptiles y anfibios. Los datos que se presentarán no muestran las especies, pero sí muestran las clases a las que pertenecería cada especie.

Nº	Clases	Nº	Clases	Nº	Clases	Nº	Clases	Nº	Clases
1	Mamífero	25	Ave	49	Ave	73	Ave	97	Ave
2	Mamífero	26	Ave	50	Ave	74	Ave	98	Ave
3	Mamífero	27	Ave	51	Ave	75	Ave	99	Ave
4	Mamífero	28	Ave	52	Ave	76	Ave	100	Ave
5	Mamífero	29	Ave	53	Ave	77	Ave	101	Anfibio
6	Mamífero	30	Ave	54	Ave	78	Ave	102	Anfibio
7	Mamífero	31	Ave	55	Ave	79	Ave	103	Anfibio
8	Mamífero	32	Ave	56	Ave	80	Ave	104	Anfibio
9	Mamífero	33	Ave	57	Ave	81	Ave	105	Anfibio
10	Mamífero	34	Ave	58	Ave	82	Ave	106	Anfibio
11	Mamífero	35	Ave	59	Ave	83	Ave	107	Reptil
12	Mamífero	36	Ave	60	Ave	84	Ave	108	Reptil
13	Ave	37	Ave	61	Ave	85	Ave	109	Reptil
14	Ave	38	Ave	62	Ave	86	Ave	110	Reptil
15	Ave	39	Ave	63	Ave	87	Ave	111	Reptil
16	Ave	40	Ave	64	Ave	88	Ave	112	Reptil
17	Ave	41	Ave	65	Ave	89	Ave	113	Reptil
18	Ave	42	Ave	66	Ave	90	Ave	114	Reptil
19	Ave	43	Ave	67	Ave	91	Ave	115	Reptil
20	Ave	44	Ave	68	Ave	92	Ave	116	Reptil
21	Ave	45	Ave	69	Ave	93	Ave	117	Reptil
22	Ave	46	Ave	70	Ave	94	Ave	118	Reptil
23	Ave	47	Ave	71	Ave	95	Ave	119	Reptil
24	Ave	48	Ave	72	Ave	96	Ave	120	Reptil

Aplicaciones de Estadística Básica

Anexo 6. Datos de la cobertura de una especie de briófito hepático llamado *Porella platyphylla* y su relación con la temperatura del aire, en tres ecosistemas, un ecosistema boscoso (Bosque), un ecosistema agrícola (Agrícola) y un ecosistema urbano (Urbano).

Nº	Árbol	Ecosis	Cober	Temp	Nº	Árbol	Ecosis	Cober	Temp
1	1	Bosque	98.1	10.2	16	6	Agrícola	70.8	16.6
2	2	Bosque	86.8	12.4	17	7	Agrícola	51.8	23.5
3	3	Bosque	88.8	10.2	18	8	Agrícola	58.4	21.8
4	4	Bosque	76.9	15.8	19	9	Agrícola	67.3	16.2
5	5	Bosque	83	14.5	20	10	Agrícola	66.1	16.6
6	6	Bosque	98.1	9.9	21	1	Urbano	44.7	24.8
7	7	Bosque	88.3	10.1	22	2	Urbano	38.9	26.4
8	8	Bosque	77.8	13.6	23	3	Urbano	26.2	30.4
9	9	Bosque	86.8	11.7	24	4	Urbano	24.6	31.2
10	10	Bosque	92.7	9.9	25	5	Urbano	38.9	28.6
11	1	Agrícola	62.9	18.2	26	6	Urbano	35	25.5
12	2	Agrícola	57.8	20.6	27	7	Urbano	40.9	25.6
13	3	Agrícola	75.1	16.9	28	8	Urbano	40.2	27.4
14	4	Agrícola	81.6	16.8	29	9	Urbano	47.4	25.2
15	5	Agrícola	49.2	22.4	30	10	Urbano	57.1	25.5

Ecosist: Ecosistema; Cober: Cobertura (%); Temp: Temperatura (°C).

Anexo 7. Datos de cobertura del briófito hepático *Porella platyphylla* que se presenta sobre la corteza de los árboles.

Árbol	Sitio	Cober	Temp	HR	CO2	Árbol	Sitio	Cober	Temp	HR	CO2
1	Bosque	98.1	10.2	92.5	316.4	16	Agricola	70.8	16.6	77.9	338.9
2	Bosque	86.8	12.4	92.4	313	17	Agricola	51.8	23.5	50.7	377.4
3	Bosque	88.8	10.2	104	304.9	18	Agricola	58.4	21.8	61.1	371.7
4	Bosque	76.9	15.8	89	342.6	19	Agricola	67.3	16.2	74.3	331.8
5	Bosque	83	14.5	75.7	322.8	20	Agricola	66.1	16.6	88	332.6
6	Bosque	98.1	9.9	91.6	300.2	21	Urbano	44.7	24.8	44.2	377.2
7	Bosque	88.3	10.1	94.8	302.2	22	Urbano	38.9	26.4	53.3	389.9
8	Bosque	77.8	13.6	85.9	324.1	23	Urbano	26.2	30.4	24.3	397
9	Bosque	86.8	11.7	97.8	316	24	Urbano	24.6	31.2	18.5	414.5
10	Bosque	92.7	9.9	92.5	307.6	25	Urbano	38.9	28.6	31.8	402.3
11	Agricola	62.9	18.2	66.1	352.7	26	Urbano	35	25.5	49.2	376.4
12	Agricola	57.8	20.6	59.4	357.6	27	Urbano	40.9	25.6	47	380.9
13	Agricola	75.1	16.9	64.8	335.1	28	Urbano	40.2	27.4	37.6	389.3
14	Agricola	81.6	16.8	68.3	347.2	29	Urbano	47.4	25.2	37.7	386.5
15	Agricola	49.2	22.4	56.7	358.6	30	Urbano	57.1	25.5	42.3	385.7

Cober: Cobertura (%); Temp: Temperatura (°C); HR: Humedad Relativa del Aire (%); CO2: Dióxido de Carbono (ppm).

Aplicaciones de Estadística Básica

Anexo 8. Datos de Oxígeno Disuelto (OD) (ppm) en 10 sitios, tomados una vez cada mes, por cinco meses.

Nº	Sitios	Meses	OD	Nº	Sitios	Meses	OD	Nº	Sitios	Meses	OD
1	1	Abril	5.6	18	8	Mayo	3.2	35	5	Julio	6.4
2	2	Abril	4.2	19	9	Mayo	5.4	36	6	Julio	5.8
3	3	Abril	3.1	20	10	Mayo	2.3	37	7	Julio	4.6
4	4	Abril	3.3	21	1	Junio	4.7	38	8	Julio	6.7
5	5	Abril	6.3	22	2	Junio	6.1	39	9	Julio	6.4
6	6	Abril	3.5	23	3	Junio	4.6	40	10	Julio	5.5
7	7	Abril	2.3	24	4	Junio	4.8	41	1	Agosto	7
8	8	Abril	3.2	25	5	Junio	5.6	42	2	Agosto	7.5
9	9	Abril	4.1	26	6	Junio	6	43	3	Agosto	7.7
10	10	Abril	8.2	27	7	Junio	3.8	44	4	Agosto	7.9
11	1	Mayo	4.5	28	8	Junio	4.8	45	5	Agosto	7.2
12	2	Mayo	5.5	29	9	Junio	5.6	46	6	Agosto	8.1
13	3	Mayo	4.6	30	10	Junio	3.9	47	7	Agosto	6.9
14	4	Mayo	4.6	31	1	Julio	5.7	48	8	Agosto	6.8
15	5	Mayo	5.4	32	2	Julio	6.5	49	9	Agosto	7.3
16	6	Mayo	5.2	33	3	Julio	5.4	50	10	Agosto	6.7
17	7	Mayo	3.6	34	4	Julio	5.8				

Anexo 9. Principales argumentos y funciones utilizadas en la sección “Estadísticas básicas en R”. Las funciones están identificadas por paréntesis de apertura y cierre (), y los argumentos con el símbolo igual (=).

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Añade objetos	R básico	paste() y paste0()	Añade un carácter a los objetos dentro de un vector. Utiliza tres argumentos, en el primero se define el vector, con el segundo se define el objeto a añadir entre comillas, con el tercero ("sep=") se define la forma (coma, punto, espacio, guion bajo, etc.) de separación entre los objetos del vector y el nuevo objeto a añadir.	paste(x, "#", sep=" ") x= Vector. #= Objeto a añadir. _ = Añadir un guion bajo entre los objetos.
Asigna nombres	R básico	colnames()	Utilizada para asignar nombres a las columnas en una matriz o tabla de datos.	colnames(x) <- c(A,B,C) x= Matriz o tabla de datos. A, B, C= Nombres.
Asigna nombres	R básico	rownames()	Utilizada para asignar nombres a las filas en una matriz o tabla de datos.	rownames(x) <- c(1,2,3) x= Matriz o tabla de datos. 1, 2, 3= Nombres.
Clasificación de objetos	R básico	class()	Clasifica el tipo de objeto o arreglo de datos.	Class(X) X= Un vector, una lista o una tabla de datos, un número o una palabra.
Combina objetos	R básico	c()	Combina objetos (números, letras, palabras, símbolos) para crear un vector.	c(1,2,3,4,5)
Combina objetos	R básico	cbind	Combina objetos por columnas.	cbind(X,Y,Z) X, Y, Z= Vectores.
Combina objetos	R básico	rbind	Combina objetos por filas.	rbind(X,Y,Z) X, Y, Z= Vectores.
Convierte objetos	R básico	as.data.frame()	Transforma vectores o matrices a formato de tabla de datos.	as.data.frame(X) X= Objeto.
Convierte objetos	R básico	as.integer()	Transforma a números enteros una variable categórica definida por letras.	as.integer(x) x= Variable categórica.
Convierte objetos	R básico	as.matrix()	Transforma vectores o tablas de datos a formato matricial.	as.matrix(X) X= Objeto.
Convierte objetos	R básico	as.vector()	Transforma tablas de datos y matrices a vectores.	as.vector(X) X= Objeto.
Crea objetos	R básico	Data.frame()	Crea tablas de datos. Por ejemplo, a partir de varios vectores.	data.frame(X,Y,Z) X, Y, Z= Vectores.
Crea objetos	R básico	select=	Con la función "subset()", selecciona las columnas de una tabla de datos.	select=c(A,B) A= Columna A. B= Columna B.
Crea objetos	R básico	subset()	Útil para extraer porciones de vectores, matrices o tabla de datos utilizando operadores de comparación (>, >=, <, <=, =, ==). Como primer argumento se indica el objeto del que se hará la extracción; luego se pueden emplear los argumentos "select=" para seleccionar columnas y "subset=" para seleccionar filas.	subset(x, select=c(A,B)) x= Objeto (tabla de datos) A= Columna A. B= Columna B.
Crea objetos	R básico	subset=	Con la función "subset()", selecciona las filas de una tabla de datos.	subset(C=C=="C1") C= Columna C. C1= Segmento C1.
Crea objetos	R básico	table()	Crea una tabla de frecuencias a partir de una variable categórica con varias categorías.	table(x) x= Variables categóricas.
Directorio/Archivos	R básico	data=	Argumento general para indicarle al programa la fuente de los datos.	data="Archivo"

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Directorio/Archivos	R básico	dir()	Muestra el contenido (archivos) en el directorio de trabajo.	dir()
Directorio/Archivos	R básico	read.csv()	Importa a R archivos con extensión de texto delimitado por comas (CSV). Con la función "file.choose()" permite abrir la ventana del explorador de Window.	read.csv(file.choose())
Directorio/Archivos	R básico	setwd()	Establece el directorio (carpeta) de trabajo.	setwd("C:/Usuario/ Mis Documentos/ Archivos R")
Directorio/Archivos	R básico	write.csv()	Exporta datos; en formato CSV (Texto Delimitado por Coma). Con el argumento "file=" se le asigna el nombre, con el argumento "row.names=TRUE" se le indica al programa que las filas tienen nombres y con el argumento "col.names=TRUE" se le indica al programa que las columnas tienen nombres.	write.csv(D, file="Nombre.csv", row.names=TRUE, col.names=TRUE) D= Variable donde se guardan los datos.
Estadística	e1071	kurtosis()	Calcula la kurtosis para un conjunto de datos. Con el argumento "na.rm=FALSE" responde a la pregunta ¿Deberían ser removidos los valores perdidos? Las opciones el argumento "type=" con 1= Algoritmo tradicional; 2= Algoritmo utilizado en los programas SAS, SPSS y Microsoft Excel; 3= Algoritmo utilizado por los programas Minitab y BMDP.	kurtosis(x, na.rm=FALSE, type=2) x= Conjunto de datos.
Estadística	e1071	skewness()	Calcula el coeficiente de asimetría. Con el argumento "na.rm=FALSE" responde a la pregunta ¿Deberían ser removidos los valores perdidos? Las opciones el argumento "type=" con 1= Algoritmo tradicional; 2= Algoritmo utilizado en los programas SAS, SPSS y Microsoft Excel; 3= Algoritmo utilizado por los programas Minitab y BMDP.	skewness(x, na.rm=FALSE, type=2) x= Conjunto de datos.
Estadística	nlme	anova()	Aplica análisis de varianza a modelos de ajuste con lm (modelos lineales) o glm (modelos lineales generalizados)	anova(x) x= modelo.
Estadística	nlme	lme()	Aplica modelos de efecto lineal mixto.	lme(m) m= Modelo.
Estadística	nortest	lillie.test()	Aplica prueba de Kolmogorov-Smirnov.	lillie.test(x) x= Conjunto de datos.
Estadística	R básico	alternative=	Utilizado para seleccionar los tipos de cola en los análisis estadísticos. Las opciones son: "less" cuando es una cola hacia la izquierda, "greater" cuando es una cola hacia la derecha y es dos cola cuando se excluye el argumento.	alternative="less"
Estadística	R básico	aov()	Aplica análisis de varianza más de dos grupos de datos.	aov(x ~ y) x= Variable numérica. y= Variable categórica.
Estadística	R básico	chisq.test()	Aplica prueba de chi cuadrado.	chisq.test(x) x= Un vector o matriz numérica.
Estadística	R básico	cor.test()	Aplica prueba de correlación. Con el argumento "method=" se define el tipo de correlación, las cuales son: "pearson", "kendall" o "spearman".	cor.test(x1, x2, method="pearson") x1= Variable 1. x2= Variable 2.
Estadística	R básico	friedman.test()	Aplica prueba de Friedman.	friedman.test(X, Y, F) X= Variable numérica. Y= Variable categórica. F= Factor.
Estadística	R básico	fun=	Define la operación estadística a utilizar en las funciones lapply, supply, tapply.	supply(x[,2:4], FUN=mean) x= Tabla de datos. mean= Media.
Estadística	R básico	kruskal.test()	Aplica prueba de Kruskal-Wallis.	kruskal.test(X ~ Y) X= Variable numérica. Y= Variable categórica.

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Estadística	R básico	lapply()	Aplica una función a las columnas de una tabla de datos. A diferencia de "sapply()" el resultado no lo presenta en forma de vector.	lapply(X[, A, B, C], FUN=mean) X= Variable donde esta guardada la tabla de datos. A, B y C= Columnas de la tabla de datos. mean= Función promedio.
Estadística	R básico	pairwise.wilcox.test()	Aplica la prueba de t de student pareada. Sus argumentos más usados son "paired=" en el que selecciona "TRUE"; si se desea una prueba para datos provenientes de medidas repetidas; "p.adj=" con el que se selecciona el método de ajuste del valor de p, las opciones son "hochberg", "holm", "bonferroni", "BH", "BY", "fdr" y "alternative=" con sus opciones: "two.sided" (por defecto), "less" y "greater".	pairwise.test(X[, Y, paired=TRUE, p.adj="holm") X= Variable numérica. Y= Variable categórica.
Estadística	R básico	prop.test()	Función general para aplicar pruebas de proporciones.	prop.test(x) x= Proporciones.
Estadística	R básico	sapply=	Aplica una función a las columnas de una tabla de datos. El resultado lo presenta en forma de vector.	sapply(X[, A, B, C], FUN=mean) X= Variable donde esta guardada la tabla de datos. A, B y C= Columnas de la tabla de datos. mean= Función promedio.
Estadística	R básico	shapiro.test()	Aplica prueba de Shapiro-Wilks.	shapiro.test(x) x= Conjunto de datos.
Estadística	R básico	summary()	Aplica estadística descriptiva a un conjunto de datos.	summary(x) x= Conjunto de datos.
Estadística	R básico	t.test()	Aplica prueba T. Uno de sus argumentos es "alternative=" con sus opciones: "two.sided" (por defecto), "less" y "greater".	t.test(x1, x2, alternative="less") x1= Primer conjunto de datos. x2= Segundo conjunto de datos.
Estadística	R básico	tapply=	Aplica una función a una variable numérica en función de una variable categórica.	tapply(N, C, FUN=mean) N= Variable numérica. C= Variable categórica. mean= Función promedio.
Estadística	R básico	tukeyhsd()	Aplica prueba de comparaciones múltiples a análisis de varianzas (ANDEVA) aplicados con la función "aov()".	tukeyhsd(ANDEVA)
Estadística	R básico	type=	Define el tipo de método utilizado para calcular la curtosis en un conjunto de datos. Las opciones son: 1= Algoritmo tradicional; 2= Algoritmo utilizado en los programas SAS, SPSS y Microsoft Excel; y 3= Algoritmo utilizado por los programas Minitab y BMDP.	type=2
Estadística	R básico	var.test()	Aplica prueba para determinar igualdad de varianzas.	var.test(G1, G2) G1= Conjunto de datos 1. G2= Conjunto de datos 2.
Estadística	R básico	which=	Selecciona entre diferentes gráficos resultantes de un análisis de regresión. Las opciones son: 1. Residuales versus valores ajustados; 2. Gráfico Q-Q; 3. Residuales estandarizados versus valores ajustados; 4. Distancia de Cook; 5. Residuales estandarizados versus valores leverage (qué tan distante están los valores de la variable independiente de una observación con relación a otras observaciones); 6. Distancia de Cook versus valores leverage.	which=3

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Estadística	R básico	<code>wilcox.test()</code>	Función general para aplicar la prueba de Wilcoxon o Mann-Whitney. Uno de sus argumentos es "alternative=" con sus opciones: "two.sided" (por defecto), "less" y "greater".	<code>wilcox.test(x1, x2, alternative="greater")</code> x1= Primer conjunto de datos. x2= Segundo conjunto de datos.
Matemática	R básico	<code>log()</code>	Calcula el logaritmo de un conjunto de números.	<code>log(x)</code> x= Números.
Matemática	R básico	<code>sum()</code>	Suma los objetos en un vector numérico.	<code>sum(x)</code> x= Un vector numérico.
Muestra datos	R básico	<code>head()</code>	Muestra las primeras seis filas de una tabla o matriz de datos.	<code>head(x)</code> x= Tabla o matriz de datos.
Muestra datos	R básico	<code>tail()</code>	Muestra las últimas seis filas de una tabla o matriz de datos.	<code>tail(x)</code> x= Tabla o matriz de datos.
Muestra datos	R básico	<code>unique()</code>	Muestra los objetos únicos en una columna de datos.	<code>unique(A)</code> A= Columna de datos.
Opciones gráficas	<code>ggplot2</code>	<code>aes()</code>	Añade la estética a los gráficos: eje X, eje Y y colores.	<code>aes(x=V1, y=V2, color=V3)</code> V1, V2, V3= Variables.
Opciones gráficas	<code>ggplot2</code>	<code>facet_grid()</code>	Crea paneles rectangulares y largos. Cuando la variable está detrás del punto separado por la tilde (~ x) los paneles aparecen ordenados por columnas. Cuando la variable está delante del punto separado por la tilde (x ~) los paneles aparecen ordenados por filas.	<code>facet_grid(~ x)</code> x= Variable categórica que servirá para definir los paneles.
Opciones gráficas	<code>ggplot2</code>	<code>facet_wrap()</code>	Crea paneles cuadrados.	<code>facet_wrap(~ x)</code> x= Variable categórica que servirá para definir los paneles.
Opciones gráficas	<code>ggplot2</code>	<code>geom_density()</code>	Define un gráfico de densidad.	<code>geom_density()</code>
Opciones gráficas	<code>ggplot2</code>	<code>geom_line()</code>	Define un gráfico de línea.	<code>geom_line()</code>
Opciones gráficas	<code>ggplot2</code>	<code>geom_violin()</code>	Define un gráfico de violín.	<code>geom_violin()</code>
Opciones gráficas	<code>ggplot2</code>	<code>ggplot()</code>	Función para la producción de un gráfico.	<code>ggplot(data=X, aes(y=A))</code> X= Conjunto de datos. A= Columna de datos.
Opciones gráficas	<code>ggplot2</code>	<code>group=</code>	Asigna agrupaciones en función "aes()" que se usa como argumento de la función "ggplot()".	<code>group=x</code> x= Una variable categórica.
Opciones gráficas	<code>ggplot2</code>	<code>levels=</code>	Con la función "factor()" establece el orden en que se deberían de presentar los paneles en un gráfico multipanel.	<code>levels=c("Ene", "Feb", "Mar")</code>
Opciones gráficas	<code>ggplot2</code>	<code>ncol=</code>	Con la función "facet_wrap()", asigna el número de columnas.	<code>facet_wrap(~ x, ncol=4)</code> x= Una variable categórica (factor)
Opciones gráficas	<code>ggplot2</code>	<code>nrow=</code>	Con la función "facet_wrap()", asigna el número de filas.	<code>facet_wrap(~ x, nrow=2)</code> x= Una variable categórica (factor)
Opciones gráficas	<code>ggplot2</code>	<code>theme_classic()</code>	Sumado a la función "ggplot()" asigna tema clásico a un gráfico, esto es cambiar el fondo gris del gráfico (por defecto) a fondo blanco y las escalas de los ejes en negro.	<code>theme_classic()</code>
Opciones gráficas	<code>ggplot2</code>	<code>title=</code>	Con la función "ggplot()", asigna el título principal al gráfico.	<code>title= "Humedad Relativa (%)"</code>
Opciones gráficas	<code>ggplot2</code>	<code>x=</code>	Le indica al programa los datos a establecerse en el eje X.	<code>x=A</code> A= Datos en la columna A.
Opciones gráficas	<code>ggplot2</code>	<code>y=</code>	Le indica al programa los datos a establecerse en el eje Y.	<code>y=B</code> B= Datos en la columna B.
Opciones gráficas	Lattice	<code>col.line=</code>	Asigna color a las líneas en un gráfico de líneas.	<code>col.line="red"</code>

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Opciones gráficas	Lattice	levels=	Con la función "factor()" establece el orden en que se deberían de presentar los paneles en un gráfico multipanel.	levels=c("Ene", "Feb", "Mar")
Opciones gráficas	Lattice	strip.custom()	Se utiliza con el argumento "strip=" dentro de la función "xyplot()". Personaliza las cajas de los nombres en un gráfico de múltiples paneles.	strip=strip.custom(bg="red")
Opciones gráficas	Lattice	strip=	Con la función "xyplot()" controla las cajas de nombres en un gráfico de múltiples paneles.	bg= Fondo de un objeto. strip=strip.custom(bg="red") bg= Fondo de un objeto.
Opciones gráficas	Lattice	xyplot()	Función para la producción de un gráfico.	xyplot(V1 ~ V2, data=D) V1, V2= Variables. D= Variable donde se guardan los datos.
Opciones gráficas	R básico	add.smooth=	Añade (TRUE) o quita (FALSE) la línea suavizada en los gráficos de punto.	add.smooth=FALSE
Opciones gráficas	R básico	alpha=	Grado de transparencia en las paletas de colores. El rango de opciones es: 0= totalmente transparente y 1= opaco.	alpha=0.6
Opciones gráficas	R básico	angle=	Indica la dirección de una barra de error añadida con la función "arrows()".	angle=90
Opciones gráficas	R básico	ann=	Dentro de la función "plot()", controla los títulos de los ejes y título principal. Cuando se establece "ann=FALSE" dichos títulos se excluyen del gráfico.	ann=FALSE
Opciones gráficas	R básico	arrows()	Añade barras de error a gráficos a modo de capa.	arrows(x, M, x, M+E, angle=90) x= Gráfico. M= Vector con medias. E= Vector con error estándar. axes=FALSE
Opciones gráficas	R básico	axes=	Dentro de la función "plot()", controla las escalas de los ejes. Cuando se establece "axes=FALSE" dichas escalas se excluyen del gráfico.	axis(side=4, line= 5)
Opciones gráficas	R básico	axis()	Añade escala a los ejes a modo de capa. Sus dos argumentos más usuales son "side=" con el que se define el lado del gráfico en donde se añadirá la escala, las opciones son 1= abajo, 2= lado izquierdo, 3= arriba, 4= lado derecho; "line=" con el que se define la línea en donde se colocará.	
Opciones gráficas	R básico	barplot()	Función general para crear gráficos de barra.	barplot(x)
Opciones gráficas	R básico	beside=	Controla posición de las barras de un gráfico de barra. Las opciones son besides=TRUE para poner una barra a la par de la otra y no escribir el argumento o escribir besides=FALSE coloca las barras de forma aplada.	x= Vector con valores de las barras. besides=TRUE
Opciones gráficas	R básico	border=	Asigna color a los bordes de un gráfico de barras.	border="red"
Opciones gráficas	R básico	bottomleft=	Con la función "legend()", inserta la leyenda de un gráfico en el extremo inferior izquierdo.	"bottomleft"
Opciones gráficas	R básico	bottomright=	Con la función "legend()", inserta la leyenda de un gráfico en el extremo inferior derecho.	"bottomright"
Opciones gráficas	R básico	box.col=	Con la función "legend()", controla el color del contorno de la caja de la leyenda.	box.col="skyblue"
Opciones gráficas	R básico	box.lty=	Con la función "legend()", controla el tipo de contorno de la caja de la leyenda. Las opciones son las mismas que las del argumento "lty=".	box.lty=2
Opciones gráficas	R básico	box.lwd=	Con la función "legend()", controla el ancho del contorno de la caja de la leyenda. El valor por defecto es 1.	box.lwd=1.5
Opciones gráficas	R básico	boxwex=	Con la función "plot()", controla el ancho de los gráficos de caja. El valor por defecto es 0.8.	boxwex=0.3
Opciones gráficas	R básico	bty	Como argumento de la función "par()", controla los bordes de un gráfico. Las opciones son: o= Aparecen todos los lados de la caja; l= Aparece solo el lado izquierdo y el de	par(bty="r")

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
			abajo; 7= Muestra el lado superior y derecho, c= Muestra el lado superior, izquierdo y abajo y l= Aparece el lado superior, derecho y abajo.	
Opciones gráficas	R básico	caption=	Con la función "plot()" aplicada a un modelo de regresión, controla el título principal.	caption="Modelo"
Opciones gráficas	R básico	cex.axis=	Controla el tamaño de los nombres y números del eje Y en un gráfico. El valor por defecto es 1.	cex.axis=1.5
Opciones gráficas	R básico	cex.lab=	Controla el tamaño de los títulos de los ejes en un gráfico. El valor por defecto es 1.	cex.lab=1.5
Opciones gráficas	R básico	cex.labels=	Controla el tamaño de los nombres asignados en la diagonal de un gráfico de relación pareada. El valor por defecto es 1.	cex.labels=1.5
Opciones gráficas	R básico	cex.main=	Controla el tamaño del título principal en un gráfico. El valor por defecto es 1.	cex.main=1.5
Opciones gráficas	R básico	cex.names=	Controla el tamaño nombres y números del eje X en un gráfico.	cex.names=1.7
Opciones gráficas	R básico	col.axis=	Asigna color a los nombres y números de los ejes.	col.axis="blue"
Opciones gráficas	R básico	col.lab=	Asigna color a los títulos de los ejes.	col.lab="green"
Opciones gráficas	R básico	col.main=	Asigna color al título principal.	col.main="yellow"
Opciones gráficas	R básico	col=	Argumento general para controlar el color. Las opciones son los colores codificados con sus nombres en inglés.	col="red"
Opciones gráficas	R básico	colors()	Al escribir y correrla en la consola de R despliega la lista de los colores.	colors()
Opciones gráficas	R básico	expression()	Utilizada para expresar expresiones con simbologías particulares, por ejemplo el símbolo de grados en la temperatura.	xlab=expression("'Temperatura" ~degree~C)
Opciones gráficas	R básico	factor()	Le indica al programa que una serie de objetos son factores (variables categóricas).	factor(x) x= Objetos par(fg="red")
Opciones gráficas	R básico	fg=	Con la función "par()" asigna color a los márgenes de un gráfico elaborado con la función "plot()".	fill="blue"
Opciones gráficas	R básico	fill=	Argumento genérico para rellenar con un color.	fontaxis=2
Opciones gráficas	R básico	font.axis=	Cambia la fuente de los nombres y números de los ejes. Las opciones son: 1= por defecto, 2= negrita, 3= cursiva (italica) y 4= negrita y cursiva	fontlab=4
Opciones gráficas	R básico	font.lab=	Cambia las fuentes de los títulos de los ejes. Las opciones son: 1= por defecto, 2= negrita, 3= cursiva (italica) y 4= negrita y cursiva	font.labels=4
Opciones gráficas	R básico	font.labels=	Controla la fuente de los nombres asignados en la diagonal de un gráfico de relación pareada. Las opciones son las mismas de "font.lab=".	font.main=3
Opciones gráficas	R básico	font.main=	Cambia la fuente del título principal. Las opciones son: 1= por defecto, 2= negrita, 3= cursiva (italica) y 4= negrita y cursiva	gap=1
Opciones gráficas	R básico	gap=	Distancia entre paneles en matriz de correlaciones gráfica.	hist(x) x= Conjunto de datos.
Opciones gráficas	R básico	hist()	Crea un histograma.	horiz=TRUE
Opciones gráficas	R básico	horiz=	Con la función "legend()", los elementos de la leyenda en secuencia horizontal.	influence.measures(x) x= Modelo de regresión.
Opciones gráficas	R básico	influence.measures()	Determina los valores de influencia en un modelo de regresión.	jitter=0.05
Opciones gráficas	R básico	jitter=	Cantidad de desagregación entre los puntos en un gráfico de puntos.	labels=X
Opciones gráficas	R básico	labels=	En las matrices de correlación pareada se utiliza para reescribir los nombres en las diagonales. También se utiliza con la función "pie()" para asignar nombre de etiquetas de cada porción del gráfico de pastel.	X= Vector con nombres.
Opciones gráficas	R básico	las=	Controla la posición de los nombre y números de los ejes X y Y. Las opciones son: 0= paralelo al axis (por defecto), 1= siempre horizontal, 2= siempre perpendicular al axis y 3 = siempre vertical.	las=2
Opciones gráficas	R básico	layout()	Crea una plantilla para diseñar un gráfico multipanel.	layout(X) X<-rbind(c(1,2),c(3,4))

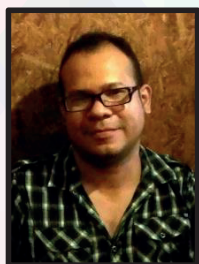
Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Opciones gráficas	R básico	legend.text=	Inserta la leyenda de un gráfico, por defecto en el lado derecho del mismo. Las opciones son legend.test=TRUE para poner la leyenda y no escribir el argumento o escribir legend.text=FALSE no poner la leyenda del gráfico.	legend.text=TRUE
Opciones gráficas	R básico	legend=length=	Con la función "legend()" establece los nombres de la leyenda.	legend=c("Sitio 1", "Sitio 2")
Opciones gráficas	R básico	line=	Con la función "arrows()" controla la longitud de la barra horizontal de la barra de error.	length=0.8
Opciones gráficas	R básico	lines()	Con la función "title()", define la línea del margen del gráfico en donde se añadirá un objeto.	line=3
Opciones gráficas	R básico		Se insertan líneas a modo de capas sobre otros gráficos.	lines(X ~ Y) X= Variable 1 Y= Variable 2 ~ = "En función de"
Opciones gráficas	R básico	locator(1)	Con la función "legend()", permite añadir la leyenda con el uso del cursor.	locator(1)
Opciones gráficas	R básico	lower.panel=	En las matrices de correlaciones pareadas se utiliza para suprimir el panel inferior seleccionando la opción "NULL".	lower.panel=NULL
Opciones gráficas	R básico	lty=	Especifica los tipos de líneas y bordes. Las opciones son: 1= sólida, 2= con guiones, 3= punteada, 4= con guiones y punteada, 5= con guiones largos y 6= con guiones largos y cortos.	lty= 2
Opciones gráficas	R básico	lwd=	Cambia el grosor de la escala del eje Y. También cambia el grosor de cualquier línea especificada. El grosor por defecto es 1.	lwd=1.5
Opciones gráficas	R básico	main=	Asigna el título principal al gráfico.	main= "Promedios de pH por sitios"
Opciones gráficas	R básico	mar=	Controla el tamaño del margen de los gráficos mediante un vector de cuatro valores que representan cada lado del gráfico. El orden inicia por el lado de abajo y continúa en el sentido de las manecillas del reloj hasta el lado derecho. Los valores por defectos se determinan con el comando "par("mar")" y los valores por defecto, por lo general, son: 5.1, 4.1, 4.1 y 2.1.	par(mar=c(5.1, 4.1, 4.1, 2.1))
Opciones gráficas	R básico	mfc=	Con la función "par()" crea una plantilla multipaneles ordenados por columnas. El número de paneles se designa por un vector.	mfc=c(2,3) # 2 filas, 3 columnas.
Opciones gráficas	R básico	mfw=	Con la función "par()" crea una plantilla multipaneles ordenados por filas. El número de paneles se designa por un vector.	mfw=c(1,4) # 1 fila, 4 columnas.
Opciones gráficas	R básico	mfp=	Con la función "par()", se le asignan tres valores, el primero controla la distancia de las leyendas de los ejes con respecto al borde del gráfico, la segunda controla las distancias de los nombres de los ejes con respecto al borde del gráfico y el tercero controla la distancia de las líneas de los ejes con respecto al borde del gráfico. Los valores por defecto se obtienen con "par("mfp")", estos son 3, 1 y 0.	par(mfp=c(3, 1, 0))
Opciones gráficas	R básico	mt=	Añade texto a modo de capa en los márgenes de un gráfico. Algunos de sus argumentos más usados son: "side=" que define el margen donde se añadirá el texto, las opciones son 1= abajo, 2= lado izquierdo, 3= arriba, 4= lado derecho; "line=" define la línea en que se añadirá el texto, puede ser desde cero a más; "adj=" define el alineamiento del texto, igual a 0 cuando el alineamiento es a la izquierda o abajo y 1 cuando es a la derecha o arriba. Otros argumentos son "cex=" "col=" y "font=".	mtext(text="Texto", side=3, line=3, adj=1)
Opciones gráficas	R básico	names=	Asigna nombre a las barras en un gráfico de barras.	names= c("Lugar 1", "Lugar 2", "Lugar 3")
Opciones gráficas	R básico	nc=	Con la función "legend()" coloca los elementos de las leyendas en columnas.	nc=4
Opciones gráficas	R básico	new=	Con la función "par()", se puede crear un nuevo gráfico sobre uno ya existente.	par(new=TRUE)

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Opciones gráficas	R básico	oneway.test()	Análisis de Varianza en el que no se asume igualdad de varianza, por lo contrario se establece el argumento "var.equal=TRUE".	oneway.test(X ~ Y) X= Variable numérica. Y= Variable categórica.
Opciones gráficas	R básico	padj=	Con la función "mtxt()", controla el número de fila en el que se añadirá un texto.	padj=3
Opciones gráficas	R básico	pairs()	Crea un gráfico de correlación múltiple pareada.	pairs(x) x= Tabla de datos con variables.
Opciones gráficas	R básico	par()	Función general para asignar parámetros a los gráficos. Por ejemplo, con el argumento "mar=" controla el tamaño de los márgenes de un gráfico; con el argumento "mfrow=" crea una plantilla para diseñar un gráfico multipanel arreglado por filas. Para mayor información escribir ?par en la consola.	par(mar=c(5,4,2))
Opciones gráficas	R básico	pch=	Cambia el tipo de símbolo en los gráficos de punto. Para conocer las opciones dirigirse al menú ayuda escribiendo ?pch en la consola de R.	pch=6
Opciones gráficas	R básico	pie()	Inserta un gráfico de pastel.	Pie(X) X= Un vector con nombres.
Opciones gráficas	R básico	plot()	Función general para crear un gráfico. Con el argumento "type=" se seleccionan los tipos de gráficos, las opciones son: p= Gráfico de puntos (por defecto); l= Gráfico de líneas; b= Gráfico de líneas y puntos con relleno blanco; c= Igual a "b" pero dejando espacio en blanco donde estaban los círculos; o= Gráfico de líneas y puntos con relleno transparente; h= Parecido a histograma, pero con líneas verticales en lugar de barras y s= Parecido a histograma, pero solamente la silueta.	plot(x, type=p)
Opciones gráficas	R básico	points()	Agrega puntos a modo de capa a un gráfico ya existente.	points(x) x= Conjunto de datos.
Opciones gráficas	R básico	pos=	Con la función "text()", ubica el texto añadido en cuatro lugares con respecto a la posición determinada por sus coordenadas X y Y: 1 debajo de la posición, 2 a la izquierda, 3 arriba y 4 a la derecha.	pos=2
Opciones gráficas	R básico	qqline()	Inserta una línea a un gráfico Q.	qqline(x)
Opciones gráficas	R básico	qqnorm()	Crea un gráfico Q.	qqnorm(x) x= Conjunto de datos.
Opciones gráficas	R básico	radius=	Con la función "pie()", controla el diámetro del gráfico de pastel. El valor máximo es 1.	radius=0.6
Opciones gráficas	R básico	side=	Con la función "mtxt()", se utiliza para indicar el lado del gráfico, las opciones son 1 que representa abajo, 2 a la izquierda, 3 arriba y 4 a la derecha.	side=1
Opciones gráficas	R básico	stripchart()	Gráfico de puntos de una dimensión.	stripchart(X ~ Y) X= Variable numérica. Y= Variable categórica.
Opciones gráficas	R básico	text()	Inserta un texto en los gráficos a modo de capa, en algún lugar asignado mediante coordenadas formadas por valores de los ejes X y Y. Con el argumento "label=" se define el texto a presentar.	text(x=1, y=80, label="Texto") x= Coordenada del eje X. y= Coordenada del eje Y.
Opciones gráficas	R básico	title()	Añade nombre a los ejes. Se puede seleccionar el eje en donde se añadirá el texto con los argumentos "xlab=" o "ylab=" y se escribe el nombre del eje entre comillas. También se controla en que línea se añadirá el texto en correspondencia con el margen del gráfico con el argumento "line=".	title(ylab="Sitios", line=4)
Opciones gráficas	R básico	title=	Con la función "legend()", asigna el título a la leyenda.	title= "Ecosistemas"
Opciones gráficas	R básico	topleft=	Con la función "legend()", inserta la leyenda de un gráfico en el extremo superior izquierdo.	"topleft"

Objeto de uso	Fuente	Argumentos y funciones	Explicación resumida	Ejemplo de su escritura
Opciones gráficas	R básico	topright=	Con la función "legend()", inserta la leyenda de un gráfico en el extremo superior derecho.	"topright"
Opciones gráficas	R básico	type=	Utilizado en la función "plot()". Define los tipos de gráficos. Las opciones son: p= Gráfico de puntos (por defecto si no se establece el tipo), l= Gráfico de líneas, b= Gráfico de líneas y puntos con relleno blanco, c= Igual a "b", pero dejando espacio en blanco donde estaban los círculos, o= Gráfico de líneas y puntos con relleno transparente, h= Parecido a histograma, pero con líneas verticales en lugar de barras y s= Parecido a histograma, pero solamente la silueta.	type=h
Opciones gráficas	R básico	upper.panel=	En las matrices de correlaciones pareadas se utiliza para suprimir el panel superior seleccionando la opción "NULL".	upper.panel=NULL
Opciones gráficas	R básico	whiskly=	Toma los mismos valores del argumento "lty=" y es útil para personalizar el tipo de línea que representa la dispersión de los datos en un gráfico de cajas.	whiskly=1
Opciones gráficas	R básico	width=	Con la función "barplot()", controla el ancho de las barras de un gráfico de barras.	width=c(3,2,5) 3, 2, 5= El ancho de tres barras.
Opciones gráficas	R básico	with()	Opción especial para crear gráficos por capa.	with(D, plot(x,y)) D= Variable donde se guardan los datos. x, y= Valores.
Opciones gráficas	R básico	xlab=	Asigna el título al eje X.	xlab= "Sitios"
Opciones gráficas	R básico	xlim=	Controla los límites del eje X.	xlim= c(A, B) A= Valor del límite inferior. B= Valor del límite superior.
Opciones gráficas	R básico	ylab=	Asigna el título al eje Y.	ylab= "Promedio pH"
Opciones gráficas	R básico	ylim=	Controla los límites del eje Y.	ylim= c(A, B) A= Valor del límite inferior. B= Valor del límite superior.
Paquetes, instalación	R básico	install.packages()	Instala paquetes, el argumento es el nombre del paquete entre comillas.	install.packages("ggplot2")
Paquetes, llamado	R básico	library()	Hace disponible un paquete, el argumento es el nombre del paquete entre comillas.	library("ggplot2")



**“Por un Desarrollo Agrario
Integral y Sostenible”**



Miguel Garmendia Zapata

Estudió Licenciatura en Biología en la Universidad Nacional Autónoma de Nicaragua, UNAN-León, Nicaragua. Con maestría en Biología Ambiental y de Bosques en la Universidad del Estado de Nueva York, Estados Unidos. Amplia experiencia en la medición de la diversidad biológica y ha impartido varios cursos referentes a estadística básica, manejo y administración de bases de datos, y análisis de datos con enfoque de biodiversidad. Actualmente forma parte del cuerpo docente, como Profesor Titular, en el Departamento de Manejo de Bosques y Ecosistemas, Facultad de Recursos Naturales y del Ambiente, Universidad Nacional Agraria.

ISBN 978-99924-1-044-8



9 789992 141044 8

*Este documento tiene la intención de servir de guía y apoyo para estudiantes de la
Universidad Nacional Agraria, Nicaragua en la aplicación de estadística
básica en Microsoft Excel y R. Esta es una obra sin fines de lucro y con objetivos educativos.*